computers

MDPI

Article

Error Pattern Discovery in Spellchecking Using Multi-Class Confusion Matrix Analysis for the Croatian Language

Gordan Gledec [1,*], Mladen Sokele [2], Marko Horvat [1] and Miljenko Mikuc [3]

[1] Department of Applied Computing, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia; marko.horvat3@fer.hr
[2] Department of Electrical Engineering, Zagreb University of Applied Sciences, Vrbik 8, HR-10000 Zagreb, Croatia; mladen.sokele@tvz.hr
[3] Department of Telecommunications, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia; miljenko.mikuc@fer.hr
* Correspondence: gordan.gledec@fer.hr

Abstract: This paper introduces a novel approach to the creation and application of confusion matrices for error pattern discovery in spellchecking for the Croatian language. The experimental dataset has been derived from a corpus of mistyped words and user corrections collected since 2008 using the Croatian spellchecker available at ispravi.me. The important role of confusion matrices in enhancing the precision of spellcheckers, particularly within the diverse linguistic context of the Croatian language, is investigated. Common causes of spelling errors, emphasizing the challenges posed by diacritic usage, have been identified and analyzed. This research contributes to the advancement of spellchecking technologies and provides a more comprehensive understanding of linguistic details, particularly in languages with diacritic-rich orthographies, like Croatian. The presented user-data-driven approach demonstrates the potential for custom spellchecking solutions, especially considering the ever-changing dynamics of language use in digital communication.

Keywords: natural language processing; spellchecking; confusion matrix; Zipf–Mandelbrot law; spelling errors; language properties

Citation: Gledec, G.; Sokele, M.; Horvat, M.; Mikuc, M. Error Pattern Discovery in Spellchecking Using Multi-Class Confusion Matrix Analysis for the Croatian Language. *Computers* **2024**, *13*, 39. https://doi.org/10.3390/computers13020039

Academic Editors: Lu Bai, Huiru Zheng and Zhibao Wang

Received: 20 December 2023
Revised: 26 January 2024
Accepted: 26 January 2024
Published: 29 January 2024

1. Introduction

Throughout written history, spelling errors have been influenced by various factors. Back in the time when people used to handwrite on paper, mistyping was result of their poor familiarity with spelling rules and orthography standards or a sign of some medical symptoms like dysgraphia. With the widespread acceptance of printing presses and typewriters, much later computers with their keyboards as standard input devices, and nowadays smartphones with virtual keyboards, a whole new set of problems opened up related to the fact that people are not perfect and simply make mistakes while using a device. In the short history of spellchecking from the late 1950s to 2020, Mitton [1] described the development of spellcheckers from dictionary lookup, affix stripping, correction, confusion sets, and edit distance to the use of gigantic databases. A comprehensive survey by Hladek et al. [2] summarizes the theoretical framework and provides an overview of the approaches developed from 1991 to 2019 related to the field of automatic spelling error detection, followed by spelling error correction.

Apart from mistyping, a common cause of spelling errors is poor knowledge of spelling rules, which applies to speakers of almost all languages. However, some languages use letters with diacritical marks (also called "diacritics") or accents that are written by users as simpler variants that are easily accessible on the virtual keyboard on the screen or do not require multiple keystrokes.

Within natural language processing, the use of confusion matrices in spellchecking plays an important role in identifying and correcting misspelled words, improving the

accuracy of language processing. Confusion matrices are particularly valuable tools in the context of spellchecking, as they provide a systematic way to analyze the performance of spellchecking algorithms by identifying the frequency of correct and incorrect correction candidates. In the field of natural language processing, confusion matrices are generally used for the descriptive statistical analysis and the visualization of words, phonemes, or tokens, but they can also be used as a starting point for exploratory analysis. In this regard, each row and each column represent a language token corpus, thereby identifying the frequency of their mutual occurrence.

The paper discusses the creation and possible application of a confusion matrix for the Croatian language derived from a dataset of mistyped words and their corrections provided by users while using the Croatian spellchecker available at https://ispravi.me/ (accessed on 31 December 2023) since 2003. The important role of confusion matrices in improving the precision of spellchecker tools, especially in the diverse linguistic context of the Croatian language, is investigated. Common causes of spelling errors are identified and analyzed, highlighting the challenges posed by the use of diacritics. The aim of the paper is to contribute to the further development of spellchecking technologies and enable a more comprehensive understanding of linguistic details, especially in languages with diacritical orthography such as Croatian.

The remainder of this paper is organized as follows: Section 2 provides insight into related research in the field of spellchecking, with particular emphasis on the use of confusion matrices, as well as on spellchecking in the Croatian language. Section 3 describes in more detail the spellchecking service that provided the data for the research and describes the language and the types of errors that users make. Section 4 describes the process of matrix creation, and Section 5 discusses each of the created matrices and highlights the implications of the obtained data. Section 6 concludes the paper and provides further insight into future work that can be based on this user data-driven confusion matrix.

2. Related Work

This section provides an exploration of the significance of confusion matrices in spellchecking, examines language technologies within the Slavic language family, and sheds light onto the language technologies and tools for the Croatian language.

2.1. Confusion Matrix

The confusion matrix is a crucial tool in natural language processing, particularly in spellchecking, as it helps in identifying and correcting misspelled words. In general, a confusion matrix lists the number of times one thing was confused with another [3]. The study of confusion matrices has been widely explored in the field of computer science, linguistics, natural language processing (NLP), and speech recognition.

In the context of NLP, Almutir and Nadeem [4] use confusion matrices to evaluate the performance of named-entity recognition systems by analyzing the discrepancies between predicted and actual entity labels. Pienaar and Snyman use them for the identification of eleven official South African languages [5]. Abandah et al. [6] use confusion matrix to correct spelling mistakes in Arabic with insufficient datasets to train the correction models.

In one study [7], the authors present a new approach to Chinese spellchecking (CSC) that prioritizes contextual similarity over traditional character similarity. The authors challenge the conventional methods of CSC; they introduce a curriculum learning framework to train models in a human-like, progressive manner that is adaptable for different CSC models. They conducted extensive experiments on the SIGHAN datasets and demonstrated superior performance over previous state-of-the-art methods, proving that focusing on contextual information significantly improves the accuracy and efficiency of spellchecking in Chinese. This research not only advances CSC but also points to a broader shift towards contextual understanding in natural language processing. In [8], the authors introduce the "Fintech Key-Phrase" dataset, a significant contribution to natural language processing in the Chinese financial high-technology sector. This dataset, comprising over

12,000 human-annotated key phrases from Chinese management discussions and analyses, addresses the lack of data resources in this domain. Key features include domain-specific content, high-quality annotations, and comprehensive evaluations, including consistency and quality assessments. The utility of the dataset is demonstrated through its integration with advanced information retrieval systems and ChatGPT for text augmentation, showing notable improvements in key-phrase extraction accuracy and coverage. Furthermore, in [9], the author compares the grammatical and semantic properties of effective constructions in English and Uzbek. The study investigates resultative structures in English, such as participles and complex objects, and compares them to similar linguistic constructs in Uzbek, with a particular emphasis on complex participles and specific suffixes that indicate resultative meanings. The study explores the differences and similarities in how these two languages use lexical, grammatical, and semantic elements to convey actions and outcomes, revealing the nuanced interaction of language units in expressing resultative meanings.

In the domain of speech recognition, Phatak et al. [10] employ confusion matrices to assess the accuracy of speech recognition systems and to identify patterns of misrecognition, aiding in the refinement of acoustic and language models. Xu et al. [11] discuss the generation of phonetic confusion matrices to enhance speech recognition performance, demonstrating the wide applicability of confusion matrices in language-related tasks.

Confusion matrices are integral to spellchecking systems, enabling the analysis of spelling correction accuracy and the identification of common spelling errors. Kernighan et al. [12] use confusion matrices to propose and sort a list of candidate corrections for misspelled words in one of the early spellcheckers named "correct," which is based on the idea of a noisy channel. They are also given considerable mention in Appendix B, of an online version of a textbook on speech and language processing by Jurafski and Martin [3].

2.2. Slavic Languages

The causes of spelling errors for the English language have been studied extensively [2]. Factors such as language interference, lack of awareness of spelling rules, and even the dissimilarity between writing systems of different languages have been highlighted as a significant cause of spelling errors. Furthermore, the use of digital tools, such as spelling software, has been explored in addressing spelling errors [13,14]. However, Slavic languages, particularly Croatian, have not been studied to such an extent [15].

The history of spellchecking in Slavic languages is deeply intertwined with the linguistic diversity and unique characteristics of these languages. The Slavic languages, traditionally divided into three distinct branches—West Slavic, South Slavic, and East Slavic [16]—have evolved over centuries, each with its own orthographic and phonetic peculiarities. The study by Golubovic and Gooskens [17] provides valuable insights into the linguistic distinctions within the Slavic language family.

The development of language technologies for Slavic languages has been a subject of interest, as highlighted by the work of Nouza et al. [18], which addresses the challenges posed by Slavic languages in automatic speech recognition (ASR) systems. The unique orthographic and morphosyntactic features of pre-modern Slavic varieties have also been the focus of research, as demonstrated by the work of Pedrazzini and Eckhoff [19], who developed a scalable Early Slavic dependency parser trained on modern language data to resemble the orthography and morphosyntax of pre-modern varieties. The linguistic diversity and historical evolution of Slavic languages have also been studied in the context of language contact and borrowing, as evidenced by the research of Adamou et al. [20], which explores borrowing and contact intensity in Slavic minority languages.

Substantial research related to n-gram systems and spellchecking has been conducted on language technologies for individual languages. For the Polish language, n-gram models were presented by Banasiak et al. [21] and Ziółko et al. [22]. Rozovskaya developed a minimally supervised model for spelling correction and evaluated its performance on datasets annotated for spelling errors in Russian [23]. Sorokin presented an algorithm for the automatic correction of spelling errors at the sentence level for Russian [24]; Richter

et al. presented a statistical text corrector tool, Korektor [25], for the Czech language; and Ramasamy et al. presented its improvements [26]. Hladek et al. [27] described a method to automatically propose and choose a spelling correction in Slovak. However, some of the problems are common to the whole language group.

The restoration of diacritic characters in Slavic languages is a significant area of research, aiming to accurately reconstruct the original orthographic forms of words. This process is particularly crucial in languages with diacritics, such as Czech, Croatian, and Polish. The restoration of diacritics involves the identification and insertion of diacritic marks to ensure the correct pronunciation and semantic interpretation of words. Research in this area encompasses various techniques, including character-based machine learning models [28]. Náplava et al. [29] propose a new architecture for diacritics restoration based on contextualized embeddings, particularly BERT, and evaluate it using 12 languages with diacritics, including Croatian. The restoration of diacritics is essential for accurate language processing and understanding in Slavic languages, and ongoing research continues to advance the development of effective diacritic restoration methods.

The research in this area has contributed to a deeper understanding of the orthographic, phonetic, and morphosyntactic features of Slavic languages, paving the way for the development of language technologies tailored to the specific needs of these languages.

2.3. Croatian Language

The Croatian language, belonging to the South Slavic branch, has a distinct orthographic system, which has influenced the development of spellchecking tools.

An innovative approach to large-scale n-gram system creation applied to the Croatian language is presented in [30]. This study highlights the efforts to develop language technologies specific to Croatian. Additionally, Šoić and Vuković [31] utilize a Croatian language network for building a solution capable of generating spoken notifications in Croatian, demonstrating the practical applications of language technologies in the Croatian context. Šantić et al. [32] describe a system for automatic diacritic restoration in Croatian texts, which combines dictionary lookup and statistical language modeling, achieving high levels of accuracy.

The advantages of online spellchecking specifically in the Croatian context, emphasizing the relevance and impact of spellchecking tools for the Croatian language, are described in [33]. This highlights the growing significance of spellchecking technologies in addressing linguistic challenges unique to Croatian.

The history of spellchecking in the Croatian language reflects concerted efforts to develop language technologies tailored to the unique linguistic characteristics of Croatian. The research in this area has contributed to the advancement of spellchecking tools and language technologies specific to Croatian, addressing the linguistic, sociocultural, and technological aspects of spellchecking in the Croatian language.

3. The Croatian Language and Common Spelling Errors

Croatia is home to the population of 4 million and is situated in Southeast Europe, on the east coast of the Adriatic Sea up to the Pannonian basin. The official language is Croatian, which belongs to the group of Slavic languages and is spoken by approximately 8 million people. It is used by Croats in Croatia and in Bosnia and Herzegovina (one of three official languages), and also in neighboring countries (in some of them as a recognized minority language). It is based on the Latin writing system, and its orthography is mostly phonetical.

Figure 1 shows a Croatian QWERTZ keyboard layout. The alphabet consists of 30 letters, 5 of them vowels. It is characterized by the usage of five letters with diacritics:

- “č”, pronounced like “ch” in the English word “checker”;
- “ć”, pronounced like “tj” in the Dutch word “Aantjes”;
- “đ”, pronounced like “Gi” in the Italian word “Giulia”;
- “š”, pronounced like “sh” in the English word “shop”;

- "ž", pronounce like "J" in the Portuguese word "Joaquim".

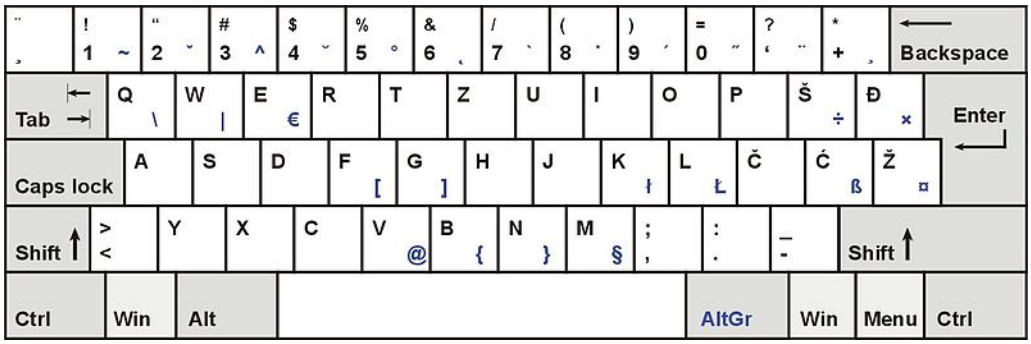

Figure 1. Standard Croatian keyboard layout [34].

Three digraphs are treated as individual letters:

- "dž", pronounced like "j" in the English word "job";
- "nj", pronounced like "ñ" in the Spanish word "señora";
- "lj", pronounced like "ll" in the Spanish word "Castilla".

The sound system uses two diphthongs—short and long "ě", which are written down as "je" and "ije". Foreign names borrow their original orthography, effectively extending the number of letters used in writing. Names from non-Latin scripts are transliterated according to Croatian rules, but in practice, often English transliteration is used. Abbreviations are written in capital letters.

The five letters with diacritics and two diphthongs are a great source of confusion for a large part of population. The three basic groups of spelling mistakes are:

- Orthography- or grammar-related errors;
- The substitution of diacritics with non-diacritics;
- Random mistyping.

3.1. Orthography- and Grammar-Related Errors

Croatian is a highly inflected language: verbs conjugate for gender, number, and tense; pronouns, nouns, adjectives, and certain numerals decline in seven cases. Nouns come in masculine, feminine, and neutral genders, and the grammatical gender of a noun affects the morphology of the surrounding adjectives, pronouns, and verbs. The abundance of orthography rules in Croatian can contribute to frequent misspellings, even among proficient speakers.

The process of orthography standardization lasted for many years, and the final orthography standard is available from the Institute for Croatian Language and Linguistics [35], but several other orthography handbooks are still in use. Orthography-related misspellings can be divided into several common types, described in the following subsections.

3.1.1. Diphthongs

In standard Croatian, the common Slavic vowel "ě" (/ie/) is reproduced as a diphthong, which is written either as "ije" or "je", but the proper variant depends on the word:

- Long /ie/—as in "tijesto" [dough] or "vijest" [news];
- Short /ie/—as in "mjesto" [place] or "vjera" [faith].

Usually, writing one instead of the other results in an easily identifiable non-word spelling error, but sometimes, adding or removing the "i" can be ambiguous. One of the notorious errors is substituting "slijedeći" for "sljedeći"—the former is used in the phrase "slijedeći zeca, završio sam u šumi" [by following the rabbit, I ended up in the forest], and the other can be used in the phrase "sljedeći dan" [next day] or "sljedeći put" [next time]. Similar examples are "svijetleći" [while one was lighting] and "svjetleći" [the one which

emits light] or "zahtjeva" [genitive of the plural of noun request] and "zahtijeva" [verb (s/he) requests] [13].

3.1.2. Diacritic Letters

Another type of common orthography error is confusing diacritic letters:

- "č" for "ć", e.g., "mač" [sward] becomes "mać", "ručak" [lunch] becomes "rućak", "četvrtak" [Thursday] becomes "ćetvrtak";
- "ć" for "č", e.g., "ćup" [cup] becomes "čup", "maćeha" [stepmother] becomes "mačeha".

As in the previous case, substitution usually leads to a non-word error, but sometimes, amusing real-word errors occur: "spavačica" [sleeping woman] vs. "spavaćica" [sleeping dress], "kuče" [small dog] vs. "kuće" [houses], "vraćati" [to return] vs. "vračati" [to cast a spell]. All those words are valid words with common usage, and detecting them as errors presents a contextual challenge.

3.1.3. Preposition "s/sa"

The third common error involves the preposition "s" or "sa" [with]. "Sa" as a preposition is used when the following word starts with "s", "z", "š", "ž", "ks", or "ps"; in all other cases, "s" is grammatically correct. Substituting one for the other is common, but the error is trivial to detect and correct.

3.1.4. Negation of Verbs

Another common error is writing negations of verbs. They are typically formed by placing the particle "ne" [not] before the verb (e.g., "ne znam" [I do not know], "ne mogu" [I cannot]), with exceptions "neću" [I will not], "nemoj" [do not], "nemam" [I do not have], and "nedostajati" [to miss]. A common error is omitting the space after the particle "ne", where instead of two words, one error word is formed (e.g., "neznam", "nemogu", etc.).

3.1.5. Future Tense

In Croatian, the future tense is formed by using the future tense of the auxiliary verb "biti" [to be], which may be "ću/ćeš/ćemo/ćete" [will], depending on the personal pronouns used. The structure is similar to the English future tense, where "will" is combined with the infinitive form of the verb (e.g., "ja ću pisati" [I will write]). If the personal pronoun is omitted, the proper form of future tense inverts the position of the verb and the auxiliary verb "ću/ćeš/ćemo/ćete" (e.g., "pisat ću"). However, many people mistakenly write the main verb in the infinitive form, with the letter "i" at the end (e.g., "pisati ću").

3.1.6. Assimilation of Consonants

The assimilation of consonants is a phonological phenomenon that occurs when adjacent consonants influence each other in terms of their pronunciation:

- Assimilation by voicing, where the voicing quality of one consonant is influenced by the voicing of a neighboring consonant that immediately follows it (e.g., "vrabac" and "vrapca" [sparrow in nominative and genitive forms], "težak" and "teška" [heavy (male and female)], "svat" and "svadba" [wedding guest and wedding]);
- Assimilation by place of articulation, which involves the modification of a consonant's place of articulation to match that of a nearby consonant (e.g., "list" and "lišće" [leaf, singular vs. plural], "grozd" and "grožđe" [grape, singular vs. plural]).

These assimilatory processes contribute to the overall fluidity and ease of pronunciation in connected speech, making language production more efficient and natural. However, the assimilation of consonants can also lead to spelling errors with users not familiar with orthography rules (e.g., writing "vrabca" instead of "vrapca").

3.2. Swapping Letters with Diacritical Marks

The second group of spelling errors stems from the fact that letters with diacritics traditionally were often substituted with their simpler variants without diacritics, especially back in the old days when keyboards and character sets did not provide support for them (e.g., ASCII character set). That substitution is still present in instant messaging and on smartphone chat apps: people write "macka" instead of "mačka" [cat], "cvjetic" instead of "cvjetić" [small flower], "skola" instead of "škola" [school], "zena" instead of "žena" [woman]. The letter "đ" is sometimes written as "d", but may also be written as "dj", although "dj" is also a legitimate digraph in Croatian—"đubre" [trash] can often be written as "dubre" or "djubre", but the word "djevojka" [girl] is a correct word that starts with "dj" and cannot be substituted with "đevojka" because that is not a valid Croatian word (but is a valid Montenegrin word).

Furthermore, words with any of the letters "c", "č", or "ć" in the same position are regular words (e.g., "placa"—genitive of colloquial for market; "plača"—genitive for the noun cry; "plaća"—salary or [s/he] pays).

In most cases, using letters without diacritics is a deliberate choice the user makes to speed up typing and by itself it does not constitute a true spelling error. Words written that way are understandable from the surrounding context, even if writing in such a way introduces real-word "errors", like when "što" [what] becomes "sto" [a hundred], "žemlja" [sort of bun] becomes "zemlja" [ground] and so on. Surely, converting words back to diacritics is a big challenge, which requires contextual spellchecking and an n-gram language model. For this task, the employed word databases enable the creation of confusion sets, since the number of such words is not too high (Table 1).

Table 1. The list of letters in the Croatian language that can be substituted with one another and cause a real-word error, with letter pairs, the number of words from the presented database that have these letters at the same position, and examples of such words.

Letters	Number of Words	Examples of Words
Č or C	1574	kolač [cake]/kolac [stick]
Ć or C	764	reći [to say]/reci [say!]
Č or Ć	579	vračati [to cast a spell]/vraćati [to return]
Č or Ć or C	48	plača [cry, gen.]/plaća [salary]/placa [market]
Š or S	2427	vaš [your]/vas [you]
Ž or Z	831	žemlja [bun]/zemlja [ground]
IJE or JE	1015	bijesni [mad]/bjesni [to act mad]
Đ or D	435	vođeno [guided]/vodeno [made of water]

3.3. Mistypings

Mistypings in writing can happen for a variety of causes, most of which are triggered by a combination of factors that affect the accuracy of keyboard input. Simple human error is one common cause, in which fingers inadvertently press the wrong keys owing to misplacement or a brief break in concentration. Fatigue and distractions can also lead to typos because fatigued or distracted typists are more likely to make mistakes.

In fast-paced typing conditions, the layout of the keyboard and the proximity of certain keys may result in inadvertent keystrokes. Furthermore, unfamiliarity with a specific keyboard layout, whether QWERTY, QWERTZ, AZERTY, or others, can add to typos, especially when users transfer between devices or regional settings. Mistypings in writing can also be caused by hearing impairment, particularly when individuals rely on auditory guidance for typing accuracy. Furthermore, individuals with hearing impairments must rely on autocorrect and spellcheck technologies to assure the accuracy of their written communication. While autocorrect and predictive text algorithms are useful, they might cause errors if they misread the intended words.

Those mistypings result either in a non-word error, which is easy to find and correct, or in a real-word error, which requires more sophisticated solutions based on understanding of the word's context.

3.4. Words from Foreign Languages and Slang

A significant share of users of the ispravi.me spellchecking service comes from Bosnia and Herzegovina, Serbia, and Montenegro, with their text written in Serbian. However, certain nuances arise from the linguistic similarities and distinctions between the Croatian and Serbian languages. Although these South Slavic languages have a shared linguistic ancestry, they have diverged over time, resulting in differences in vocabulary, spelling, and grammatical subtleties. Croatian spellcheckers may not reliably identify Serbian-specific vocabulary, phrases, or grammatical patterns, which could result in incorrect evaluations or omissions while reviewing for errors.

Such problems arise in diphthong use—the short and long /ie/ are in Croatian written as "je" and "ije", while in the Serbian language, both are written as "e" (e.g., in Croatian we write "rijeka" [river], "mlijeko" [milk], "pjevati" [to sing]; in Serbian, those words become "reka", "mleko" and "pevati". In most cases, the usage of a Serbian word will be marked as a spelling error, but sometimes, it may cause a real-word error (e.g., Croatian: "ljeti" [during the summer], Serbian: "leti"—in Croatian it means [he/she/it flies]).

The modern Croatian language has also experienced the increasing influence of English words on various domains, particularly in the realms of technology, business, and popular culture. As Croatia is connected globally and engages in international exchanges, English terms often find their way into everyday conversations and written texts. This infusion of English vocabulary poses a challenge for spellchecking in Croatian texts and extends possible spelling errors.

3.5. Ispravi.me—Croatian Online Spellchecker

Almost thirty years ago, in March 1994, the spellchecker for the Croatian language was introduced as an online email service, starting from a small corpus of 100,000 words derived from a Croatian–English dictionary and a corpus of words in English borrowed from the Unix spelling program. In 2003, email service was transferred to the World Wide Web, and the usage of the service has grown ever since. During the email phase, the service only listed suspicious words, without offering corrections. The suggested corrections were added as the service migrated to the web. Each time users chose the proper correction candidate, the pair "error word $\rightarrow$ correct word" was logged on the server. That gave us a huge dataset, published in [36].

The architecture of the Croatian Academic Spelling Checker (Croatian: "Hrvatski akademski spelling checker", abbreviated as "Hascheck" and pronounced as "Hašek", as it was known for more than 20 years) is extensively described in [37].

Briefly, as the text arrives for analysis, the Extractor block extracts valid tokens and removes them from further processing. Non-recognized tokens are then passed to the Classifier, which forwards them to the Guesser and the Corrector, which consult the Dictionary and suggest corrections in the final report sent to the user. Learning is performed offline and is supervised by an administrator. Learning is based on the data collected during usage (statistics, logs, input text, and reports). As the result of the learning process, the dictionary is updated under human supervision, thus improving the spellchecker's functionality.

Spellchecking is not based on a static corpus; it is based on live traffic, created by real people of all sorts of professions—journalists, scientists, translators, writers, lawyers—but also by regular people who just use it to spellcheck their personal correspondence. Unlike static newspaper or book corpus, ispravi.me's growing crowdsourced database includes modern words, slang, abbreviations, named entities, etc.

The dictionary is organized in three word-list files: word types, name types, and English types. The initial word type file was derived back in the 1970s from the English–

Croatian Lexicographic Corpus (ECLC), which produced 100,000 words that may occur written in small letters only, with an initial capital letter at the start of a sentence, or in capital letters only. In 30 years, the word type file grew to 1,108,164 tokens as of December 2023.

The left-hand side of the ECLC was used to produce 70,528 different English word types. The reasoning for the inclusion of English words is this: as the modern lingua franca, English, often comes mixed with Croatian words. Words that are shared between languages were removed from the English types on file. It is the only dictionary file that has not changed at all since it was created.

The name type file contains all the case-sensitive elements of writing: proper and other names, abbreviations, and acronyms, as well as names with the unusual use of small and capital letters, like LaTeX. It also contains words from foreign languages that appear in Croatian writing in their original orthography. The file started empty, but over the course of learning, it increased to 1,088,606 name types as of December 2023.

The service is available online at https://ispravi.me/ (accessed on 31 December 2023) [correct.me], and as of December 2023, according to the collected server statistics and Google Analytics data, it serves almost 12,000 user sessions per day. From 2003 until December 2023, Hascheck processed almost 62 million texts which form a corpus of 15.8 gigatokens (Gtokens). The service registered usage by almost 2 million IP addresses.

The ispravi.me server keeps track of spelling errors that were found in received texts and suggestions sent to the user, text statistics (number of different classes of errors, number of words and characters in incoming texts), and valid words selected by users from the list of suggested words. Incoming texts are subjected to n-gram analysis, which over the years has resulted in an n-gram system for Croatian language [38]. After n-gram processing, incoming texts are removed from the server for reasons of maintaining user privacy.

In [36] the authors presented an extensive dataset containing a total of 33,382,330 entries of the form "error word → correct word" collected between December 2008 and March 2023 compiled from the contributions of nearly 900,000 users of ispravi.me, the most popular Croatian online spellchecking service. In this huge dataset, the authors identified 5,584,226 unique "error word → correct word" pairs. In total, 5,296,266 unique words were misspelled, which the authors corrected to a total of 1,530,329 words. The authors use this dataset as a foundation for the creation of a letter-level confusion matrix for the Croatian language. Every record of the dataset includes the record date, the ID of the request, the error word, the correct word chosen by user, and the Damerau–Levenshtein edit distance. A sample of the dataset is given in Table 2.

Table 2. A sample of the dataset of misspelled words and their corrections.

Date	ID	Error Word	Correct Word	Edit Dist.
1 January 2023	1CF4581A-8A08-11ED-B704-EE0D37D1B59E	pdštampala	odštampala	1
1 January 2023	7A46FEB2-89A6-11ED-B68F-D29936D1B59E	neradimo	ne radimo	1
1 January 2023	18E119D6-B70C-11EC-B7C7-DD6037D1B59E	ispresjecanom	ispresijecanom	1
1 January 2023	18E119D6-B70C-11EC-B7C7-DD6037D1B59E	sa kamenim	s kamenim	1
1 January 2023	18E119D6-B70C-11EC-B7C7-DD6037D1B59E	sa minaretima	s minaretima	1
1 January 2023	18E119D6-B70C-11EC-B7C7-DD6037D1B59E	stanvništvo	stanovništvo	1
1 January 2023	04C11928-70B2-11ED-9283-780836D1B59E	dezurne	dežurne	1
1 January 2023	04C11928-70B2-11ED-9283-780836D1B59E	preumjerim	preusmjerim	1
1 January 2023	8E1D880A-3E4E-11ED-94CF-04A637D1B59E	cklusa	ciklusa	1
1 January 2023	8E1D880A-3E4E-11ED-94CF-04A637D1B59E	će doživjet	će doživjeti	1
1 January 2023	8E1D880A-3E4E-11ED-94CF-04A637D1B59E	prouzročiti će	prouzročit će	1

4. Confusion Matrix

A vital tool in natural language processing, especially for spellchecking, is the confusion matrix, which aids in locating and correcting misspelled words by providing probabilities that one word will be transformed into another.

In order to measure how close the error word is to the correct word (edit distance), the Damerau–Levenshtein metric is used to identify the minimum number of insertions, deletions, substitutions, or transpositions of a single character needed to transform the error word into a correct one [39]. If the correct word can be generated using only one transformation, the edit distance between the error word and the correct word is 1. If two basic transformations are required, then the edit distance is 2, and this pattern continues accordingly.

The confusion matrices will provide counts, relative frequencies, or probabilities indicating that a given spelling mistake happened at a given location in the word. For example, a substitution matrix for Croatian will be a square matrix of 30 $\times$ 30 letters, which represents the number of times one letter was incorrectly used instead of another. A transposition matrix will tell us how many times two letters were erroneously swapped.

The relative frequencies of inserting or deleting a specific letter can depend on either the preceding or the following character. Both approaches are utilized and will be detailed in the subsequent section. In order to calculate the relative frequency for each edit, a confusion matrix is required that records the counts of these errors.

4.1. Creation of Confusion Matrices

To create the confusion matrices, a subset of the ispravi.me dataset for the period 2008–2016 was used, which contained a total of 1,011,307 unique pairs of "error word $\rightarrow$ correct word". Those pairs appeared 3,489,162 times in the texts users corrected through the ispravi.me web service interface.

" During the process of matrix creation, the letters from the Croatian alphabet were converted to lowercase. The letters "dž", "lj", and "nj" were omitted from the analysis because they are digraphs and always written as two letters (even though the UTF-8 character set supports them as one letter, that option is seldom used). Restricting the matrix to the Croatian alphabet, the English letters "q", "w", "x", and "y", which are not part of the Croatian alphabet, were omitted, even though they appear in English words and in the named entities database.

After excluding words containing letters that do not belong to Croatian alphabet, the entries in the form "error word $\rightarrow$ correct word" where the Damerau–Levenshtein edit distance (the selected measure of choice) between the error and correct word was equal to 1 were extracted. That left a corpus of 824,959 unique pairs that contained 3,009,996 transformations that were subsequently further analyzed.

4.2. Types of Matrices

The task that followed was to parse the errors and create the matrices. Iterating over the list of all pairs with edit distance 1, it was determined which of the four types of edits—insertions, deletions, substitution, or transpositions—occurred using the following Algorithm 1:

Algorithm 1: Determining the type of an edit for a given pair of [error, correct]

```
for each pair [error, correct] do
    if DL_edit_distance(error, correct) = 1 then
        if length (error) > length (correct) then
            return {insertion}
        elseif length(error) < length(correct) then
            return {deletion}
        elseif diff (error, correct) = 1 then
            return {substitution}
        else
            return {transposition}
        end if
    end if
end for
```

Table 3 summarizes the types of identified transformations. Among all errors, substitution dominates: if sorted by descending frequency, in the first 10 errors, 6 are the result of substitution, 3 of insertion, and 1 of deletion.

Table 3. Number and share of detected Damerau–Levenshtein edit distance 1 transformations.

Transformation	Count	%
Insertion	663,014	22.03
Deletion	893,562	29.69
Substitution	1,154,149	38.34
Transposition	299,271	9.94
Total	3,009,996	100.00

4.3. Conditioning Insertion and Deletion on Both the Previous and Following Letters

Although similar to research results from four confusion matrices (e.g., [12]), one for each transformation type, due to the nature of the most common errors in Croatian, two subvariants of both deletions and insertions (conditioning on the previous and the following letter) were used. More precisely, a total six confusion matrices were created:

1. insertionCondOnFollowing—letter Y inserted in front of letter X (X → YX);
2. insertionCondOnPrevious—Y inserted after X (X → XY);
3. deletionCondOnFollowing—Y deleted in front of X (YX → X);
4. deletionCondOnPrevious—Y deleted after X (XY → X);
5. Substitution—Y substituted for X (X → Y);
6. Transposition—switching adjacent X and Y (XY → YX).

The reason for the choice of six matrices is explained in Section 3: common errors are inserting "i" before "j", deleting "i" before "j", and inserting or deleting "a" after "s". So, conditioning on both the previous and following character in insertions and deletions is appropriate:

- insertionCondOnFollowing is convenient when it is necessary to track where "i" was mistyped before "j"; otherwise, those errors would be spread to all the cases where "i" was added after any other letter in the insertionCondOnPrevious;
- insertionCondOnPrevious is convenient to track errors where "sa" was wrongly used instead of "s" [with]; otherwise, the insertions of "a" before space characters in insertionCondOnFollowing must be tracked;
- deletionCondOnFollowing is convenient to track where "i" was mistakenly deleted before "j"; otherwise, those errors would be spread to all the cases where "i" was deleted after any other letter in deletionCondOnPrevious;
- deletionCondOnPrevious is convenient to track errors where "sa" should be a proper preposition instead of "s", as one would need to track "a" missing before the space,

which would include cases where "na" [on] was misspelled as "n", "za" [for] as "z", "da" [yes] as "d, "ja" [I] as "j", etc. in deletionCondOnFollowing.

Table 4 gives clear insight into the most common orthography-related mistakes explained earlier in the paper: writing "je" instead of "ije", converting diacritics, and the wrong usage of "s" and "sa" prepositions. The ten most common errors account for 48.92% of all errors in the presented dataset.

Table 4. Ten most common transformations in the studied dataset.

Rank	Transformation	Correct Letter	Wrong Letter	Count
1	Deletion	j	i	212,065
2	Substitution	ć	č	188,320
3	Substitution	č	c	166,190
4	insertionCondOnPrevious	s	a	163,015
5 [1]	insertionCondOnFollowing	<space>	a	160,750
6	Substitution	š	s	151,514
7	insertionCondOnFollowing	j	i	149,241
8	Substitution	ž	z	106,936
8	Substitution	ć	c	103,852
10	Substitution	č	c	103,615

[1] Transformations #4 and #5 both reflect the "s/sa" error.

4.4. Space and Word Boundaries

Apart from the 30 letters of Croatian alphabet, the insertion and deletion matrices contain one more column and two more rows. The space character is present in both a row and a column (represented as "␣"), since the dataset contained a number of spelling errors containing two-word expressions:

- "sa tobom" → "s tobom" [with you]—"a" deleted in front of a space;
- "bi smo" → "bismo" [(we) would]—space inserted after "i" or before "s";
- "neznam" → "ne znam" [I do not know]—space deleted after "e" or before "z";
- "oprostiti ću" → "oprostit ću" [I will forgive you]—"i" inserted before space", etc.

A space in the error word is the result of the ispravi.me spellchecker targeting the exact type of the common error. Had the spellchecking been restricted to just one word, it would not be possible to find this mistake. Explanations for both errors are given in Section 3.1.

The word boundary (represented as "@", as in [12], meaning the beginning or the end of a word) is in the last row because the character can be inserted or deleted at the beginning or the end of the word:

- Insertion: "adodati" or "dodatia" → "dodati" [to add];
- Deletion: "apsodija" or "rapsodij" → "rapsodija" [rhapsody];
- The option existed to remove those two characters to maintain matrices at 30 × 30 letters, but this could lead to inconsistencies since the total count of errors would not be the same when conditioned on the previous or the following letter.

4.5. Content of the Confusion Matrices

Using a subset of data from the authors' extensive dataset [36], three matrices for each type of error with the following values were created:

1. Number of times the error occurred;
2. Relative frequencies of an error on a given letter;
3. Relative frequencies of an error with respect to the whole analyzed subset.

The data from all three matrices are already available online as a result of the authors' previous study [40]. In each of the published matrices, by selecting the value in the row/column intersection, examples from the dataset for each type of error may be provided.

Regarding the terms used in the paper for the description of frequencies, it is important to emphasize that the term relative frequency was used instead of probability. Also, for obtained values in confusion matrices, the term relative frequencies was used instead of probabilities. These two concepts are related, but they have some subtle differences. Both represent measures used to describe the likelihood of events; however, the relative frequency is based on observed data from observations, while probability is a theoretical measurement of the likelihood of an event occurring. Since the presented research is based on observed data, the correct term, relative frequency, was used instead of probability.

5. Discussion

In the following section, numerical tables with a heatmap-like visualization of a confusion matrix for each type of edit are presented. In all six confusion matrices shown below, the rows represent the letter X, the columns represent the letter Y, and the number at the intersection represents the relative frequencies (RFs) of error_{XY} and is displayed as $-\log_{10}(\text{RF}(\text{error}_{XY}))$ for the given type of spelling error, rounded to two decimal places. The logarithmic scale is used in this paper due to the limited space, since the original values that are available online [40] contain too many decimal places to be presented.

A log scale with heatmap-like visualization offers a good insight into our conclusions about error patterns in the Croatian language. However, when using the matrix, we strongly recommend using the data availabe online, as relative frequency values are significantly more precise than the log-scale values presented in this paper.

The matrices should be read as follows. For example, in the *insertionCondOnPrevious*, at the intersection of row "j" and column "i" is number 0.65, which means that the relative frequency of "i" being mistakenly added after "j" is $10^{-0.65}$, which amounts to 0.22387. In [40], available online, the value at that intersection is presented more precisely as 0.225095 or 22.5095%.

The lower the value in the matrix, the greater the relative frequency of this error. In each table, the values are colored to visualize the most frequent errors: the color of each cell can gradually change from green (high cell values—low relative frequency) to red (low cell values—high relative frequency).

Rows and columns for digraphs "dž", "lj", and "nj" are omitted from all matrices. Space and word boundary are omitted from the substitution and transposition matrices since they have no significant associated counts.

5.1. "insertionCondOnFollowing" Matrix

Table 5 presents the relative frequencies of errors where X was mistyped as YX (X → YX). The two most frequent errors, accounting for almost a half of all insertion errors, are:

- Wrong usage of the preposition "s/sa"—recorded as "a" added before space, as explained in Section 3.1.3.—representing 24.25% of all insertion errors (in the matrix, it is represented as the value 0.62 at the intersection of row "␣" and column "a". As suggested, we refer the reader to our online data, and at that intersection is the value 0.242453, which is the relative frequency of that type of error ($-\log_{10}$ 0.24243 is 0.615372437, rounded to 0.62 in this table). Examples of such mistakes are also available in [40] by clicking on the cell value. Some of the notable examples include ("sa tim" instead of "s tim" [with that], or "sa drugim" instead of "s drugim" [with another].
- Inserting "i" in front of "j", as explained in 3.1.1., with over 100,000 occurrences of that type (22.51% of all insertion errors), the most common being writing "riješenje" instead of "rješenje" [solution] (intersection of row "j" and column "i").

Table 5. *"insertionCondOnFollowing"*—relative frequencies of errors with edit distance 1 where letter Y was mistyped before X, i.e., X was mistyped as YX (X → YX).

	a	b	c	č	ć	d	đ	e	f	g	h	i	j	k	l	m	n	o	p	r	s	š	t	u	v	z	ž	␣
a		3.49	3.40	3.26	3.83	3.08	4.52	2.51	3.41	3.48	3.27	2.62	2.36	2.95	2.64	3.12	2.64	2.60	3.30	2.39	2.93	3.81	2.54	3.13	2.98	3.10	4.21	5.82
b	3.13		4.98	5.82	5.34	2.93	5.12	3.19	5.52	4.36	4.98	3.45	4.29	4.71	3.99	3.87	3.62	3.03	3.73	3.88	3.49		4.54	3.57	3.52	4.15	4.98	
c	2.77	5.34		4.82	5.52	4.11		3.42	4.65	4.98	4.92	2.75	3.96	3.58	4.11	4.65	3.49	3.35	4.25	3.60	3.13	5.34	3.40	3.91	3.36	4.78	5.82	5.52
č	3.19	5.34	4.16		3.35	4.54	5.52	3.79		5.04	5.52	3.35	3.52	4.02	2.88	5.22	3.70	3.75	4.02	3.94	4.13	4.29	4.06	3.64	4.50	5.04	5.04	
ć	4.07		4.23	3.28		4.87		3.88		5.34		3.72	4.87	5.82	4.78	5.82	4.52	4.20	4.82	3.94	5.34	4.42	4.25	4.27	5.22	4.57	3.80	
d	2.74	4.34	4.36	5.82	5.04			3.00	3.88	4.11	4.87	2.96	3.68	3.78	3.85	4.12	2.65	2.83	3.66	3.32	3.17		3.84	3.45	3.89	3.89	5.22	5.82
đ	4.33	5.82				3.72		4.44		5.52		4.54	4.87	5.82		5.82	4.32	4.54	4.82	4.98	5.12	4.11		4.78	5.82	4.87	4.39	
e	2.91	3.63	3.44	3.51	3.42	3.34	4.02		3.39	3.76	3.17	2.54	2.14	3.09	2.90	3.19	2.79	3.03	3.12	2.43	3.14	3.65	2.76	3.14	3.47	3.29	3.94	
f	3.91	5.34	5.82			4.10		3.51		4.25	5.34	3.99	5.52	4.57	4.08	4.74	4.50	3.46	4.87	4.21	4.25		4.82	4.41	4.65	5.34		
g	3.09	4.29	5.52	5.82		3.53		3.36	3.80		3.67	3.03	4.19	4.03	3.82	4.12	3.45	2.66	3.74	3.18	3.93	5.52	3.88	3.60	4.44	4.10	5.12	
h	3.60	5.22	3.60			4.48		4.11	5.52	3.69		3.34	3.67	4.42	4.33	4.98	4.33	3.30	4.98	4.23	4.04	5.34	4.02	4.11	4.68	4.29		5.82
i	2.43	3.40	3.25	3.29	3.71	3.37	4.44	2.40	3.65	3.94	3.48		2.38	3.09	2.47	2.99	2.18	2.44	3.27	2.52	2.85	3.73	2.38	2.35	3.11	3.17	4.23	5.12
j	2.78	4.10	4.19	3.45	5.52	3.55	5.34	2.81	5.34	4.52	3.33	0.65		3.20	2.62	3.39	2.85	2.79	3.56	3.74	3.55	5.12	3.75	3.17	3.51	4.34	5.82	
k	2.37	4.98	3.50	3.98	4.50	3.75	4.82	3.00	5.22	4.42	3.97	2.80	3.07		3.20	3.52	3.11	2.99	3.54	3.40	3.00	4.22	2.93	3.06	4.28	4.28	5.52	5.52
l	2.27	3.96	4.62	3.07	5.04	3.60	4.98	2.79	4.42	3.56	4.22	2.52	3.34	2.76		4.16	3.51	2.57	3.35	3.47	3.15	4.22	3.10	3.19	3.47	3.83	5.22	5.34
m	2.59	4.34	4.42	5.82	5.82	3.69		2.72	4.68	4.36	4.16	2.69	3.14	3.51	3.65		3.02	2.76	3.53	3.00	3.05	5.82	3.82	3.23	4.02	4.15	5.52	
n	2.00	2.72	3.85	3.78	4.21	3.15	4.44	2.14	4.42	4.02	3.96	2.17	3.08	3.70	3.05	3.02		2.42	3.45	2.91	2.84	4.04	2.40	2.93	3.21	3.51	4.41	5.34
o	2.34	3.46	3.89	3.85	5.82	3.31	5.82	2.61	3.91	3.63	3.56	1.99	2.27	2.87	2.84	3.01	2.48		2.61	2.50	3.09	3.40	3.11	3.13	3.13	3.70	5.34	5.82
p	2.61	4.24	4.50	4.78	5.52	3.88		2.89		4.98	4.57	3.12	4.00	4.03	3.89	3.37	3.88	2.50		3.50	2.87	2.82	3.53	2.97	4.62	4.28	5.34	
r	2.32	3.94	4.08	4.87	5.82	3.11		2.52	3.30	3.60	3.63	2.71	3.84	3.64	3.83	4.15	3.49	2.30	2.84		3.12	3.55	2.19	3.19	3.55	2.85	5.82	
s	2.55	4.21	4.27	5.52		2.99		2.63	4.68	4.23	4.48	2.60	3.26	3.58	3.52	3.94	3.07	2.65	3.29	3.22		4.92	2.81	2.86	3.79	3.53		2.54
š	3.27	5.82	5.52	4.82	4.87	5.04	3.36	3.49	5.52	4.13	5.52	3.14	4.30	4.19	4.52	4.92	3.91	3.61	2.96	4.02	3.13		3.95	3.61	4.44	4.24	4.98	
t	2.21	4.44	3.80	4.44	5.22	3.21	4.46	2.58	4.18	3.74	4.04	2.27	3.40	3.21	3.35	3.76	2.64	2.62	2.93	2.71	2.43	3.83		2.98	4.16	3.27	4.68	5.82
u	2.61	3.95	4.25	3.98	4.28	3.54	5.22	2.82	4.22	4.02	3.46	2.72	2.71	3.29	2.98	3.50	3.20	2.74	3.66	3.06	3.27	4.11	3.07		3.67	3.03	4.71	5.82
v	1.98	2.64	3.13	5.82	5.82	3.48		2.77	4.41	3.95	4.21	2.87	3.53	3.66	3.70	4.36	3.56	2.27	3.53	3.07	3.19	4.92	3.35	3.27		3.69	5.34	
z	3.27	4.65	4.65	5.22		3.55		3.51		4.82	4.44	2.98	4.07	4.57	4.12	4.59	3.45	3.10	4.02	3.58	3.31	5.52	3.11	3.20	4.50		5.82	
ž	3.18	5.34	5.12	5.52	3.58	4.20	3.98	3.75	5.12	4.74		3.70	4.11	5.52	4.92	5.34	4.28	4.16	4.44	4.21	4.74	5.04	4.78	4.01	5.12	3.75		
␣	0.62											1.57																
@	2.17	3.39	3.40	3.89	4.11	2.76	4.68	2.38	3.84	3.23	2.97	2.10	2.69	3.01	2.52	2.31	2.42	2.35	3.04	2.88	2.33	3.96	2.40	2.41	3.52	3.48	4.07	3.19

Note: Rows and columns for letters with empty cells are omitted from the table.

5.2. *"insertionCondOnPrevious" Matrix*

Table 6 presents the relative frequencies of errors where X was mistyped as XY (X → XY). Here, the error of writing "sa" instead of "s" is the most frequent, accounting for almost a quarter of all insertion errors. The only error that exceeds the 5% share is adding "i" after "v", which is due to the "ije/je" subcase (e.g., "uvijet" instead of "uvjet" [condition], "savijet" instead of "savjet" [advice]). Other notable mentions include adding "i" after "r", "l", "t", "m", "c", "p", and "d" (intersection of column "i" and rows "r", "l", "t", "m", "c", "p", and "d". This illustrates why conditioning on the previous character makes more sense for that type of error.

It is worth considering the differences in treating insertion errors when X and Y are the same letter (duplication), e.g., writing "zebbra" instead of "zebra". When the correct and wrong words are matched, the first occurrence of a duplicate letter is considered correct (X); the second is considered an error (Y). So, in the insertionCondOnFollowing matrix, the second letter is considered the wrong letter inserted before the next character; therefore, the main diagonal of Table 4 is empty.

In the insertionCondOnPrevious matrix, the duplicate letter inserted after the correct one produces X → XX, so the main diagonal has values. The dataset shows that the most duplicated letter is "i", with word "niije" written instead of "nije" [not] most often.

5.3. *"deletionCondOnFollowing" Matrix*

Table 7 presents the relative frequencies of errors where YX was mistyped as X (YX → X). The error of deleting an "i" in front of "je" is the most frequent of this type of error (intersection of row "j" and column "i"). The most common errors of this type are writing "uvjek" instead of "uvijek" [always] and "promjeniti" instead of "promijeniti" [to change].

Table 6. *"insertionCondOnPrevious"*—relative frequencies of errors where X was mistyped as XY (X → XY).

	a	b	c	č	ć	d	đ	e	f	g	h	i	j	k	l	m	n	o	p	r	s	š	t	u	v	z	ž	␣
a	2.49	2.94	3.35	3.46	3.55	3.05	3.84	3.36	4.24	3.78	3.53	2.60	2.80	2.74	2.60	2.84	2.50	2.75	3.37	2.94	2.44	3.69	2.47	2.90	3.14	3.27	4.21	3.92
b	2.95	3.28	5.34	4.36		4.42		3.04	4.87	4.37	4.50	2.38	4.02	3.80	3.47	4.32	3.45	2.87	4.41	3.28	4.52		3.90	3.71	3.17	4.82	5.82	
c	3.50	5.34	3.21	5.82		4.48		3.48	4.48	5.82	3.52	1.77	4.04	3.81	4.23	4.25	4.09	3.59	4.78	4.52	4.22		4.20	3.81	3.23	5.04		
č	2.96	4.65	4.65	3.76	3.59	5.52		3.20			5.82	2.39	3.76	3.85	3.05	4.65	3.47	4.32	4.30	4.82	5.22		4.32	3.52	4.74	4.92		
ć	4.00	5.12	4.54	3.40	3.97			3.28		5.52		3.71	4.27	4.92	5.04	4.87	4.19	4.44		5.52	4.78	4.57	4.50	3.69	5.52	5.82	4.16	5.82
d	2.57	3.69	4.46	4.62	5.04	2.95	5.04	2.76	3.18	4.39	4.33	1.84	3.15	3.97	3.45	4.32	2.67	2.59	4.09	3.22	3.25	5.82	3.47	3.31	3.78	4.11	4.74	4.92
đ	4.28	5.52		5.82		4.48		4.30		5.22		4.46	3.92				5.22	5.34		5.22		3.78	5.12	5.12	5.82	5.34	4.57	
e	2.66	3.24	3.55	3.48	4.17	2.87	4.06	2.60	4.15	3.75	3.85	2.64	2.85	3.20	3.22	3.03	2.49	2.93	3.48	2.35	2.65	4.00	2.73	2.95	3.72	3.59	4.06	3.99
f	4.24		5.52	5.34		4.92		3.77	3.21	4.74	5.52	3.42	5.82	4.65	4.82	5.52	4.59	3.55	5.04	3.72	3.87		4.13	4.05	5.34			
g	2.96	4.37	5.82	4.74		4.28	5.52	3.28	3.64	3.26	3.27	3.28	4.34	4.23	3.56	4.37	3.73	3.03	4.19	3.53	3.90	5.82	3.87	3.57	4.14	4.32	5.82	4.54
h	3.12	4.39	4.98	5.82		5.22		3.47	5.82	4.20	3.66	3.37	3.81	4.71	4.82	4.52	4.16	3.96	4.68	4.04	4.57		3.24	4.29	4.52	4.33	5.34	4.82
i	2.75	3.53	3.28	3.78	3.85	3.29	4.08	2.95	4.32	3.66	3.29	2.17	2.12	3.09	2.79	2.95	2.47	2.06	3.13	3.12	2.70	4.05	2.36	2.67	3.36	3.37	4.27	2.53
j	2.55	4.41	4.36	4.82	5.52	3.92		2.33	5.82	4.92	3.32	2.48	2.87	2.95	3.56	3.86	3.08	3.34	4.44	3.91	3.46	4.62	4.03	2.97	3.94	4.44	5.12	5.82
k	1.95	4.23	4.09	4.87	5.52	4.11		3.06	5.22	4.74	4.87	2.82	3.27	2.97	2.59	3.83	3.78	2.39	3.98	3.38	3.50		3.14	3.19	3.72	4.54	5.12	4.92
l	2.59	4.57	4.48	2.85	5.34	4.09		2.92	5.52	4.41	4.10	1.39	2.61	3.11	2.29	4.23	3.14	2.84	4.07	3.62	3.63	5.82	3.88	3.31	4.44	4.21	5.52	5.52
m	2.43	4.65	4.52	4.92	5.22	3.78		2.59	5.52	4.74	4.82	1.55	3.13	3.35	3.62	2.58	2.84	2.84	3.82	3.87	3.41	4.78	3.96	3.45	4.74	4.98	5.34	4.22
n	2.19	3.24	3.58	5.04	5.52	3.17	5.34	2.46	4.42	3.55	3.91	2.15	2.06	3.45	3.81	2.86	2.52	2.69	3.74	3.89	2.91	5.52	2.81	3.08	3.94	3.78		4.92
o	2.69	2.88	3.51	4.00	4.14	2.47	3.95	2.89	3.92	3.34	3.64	2.37	3.01	3.34	3.07	2.89	2.84	2.29	2.36	2.98	2.48	3.88	2.72	2.79	3.20	3.69	4.41	4.05
p	2.81	4.87	4.48	3.83	5.12	4.33		2.95	4.87	4.98	4.52	1.83	4.13	3.83	3.19	4.46	3.83	2.37	2.69	2.56	3.49	2.97	3.04	3.58	4.57	4.39		
r	2.31	4.02	3.88	4.52	4.44	3.43	4.11	2.17	3.92	3.96	4.23	1.35	3.03	3.66	3.70	3.61	3.11	2.68	3.54	2.60	3.09	4.57	2.56	3.01	3.61	3.66	4.68	5.52
s	0.61	3.96	3.48	4.30		3.13		2.89	4.68	4.08	3.50	2.00	3.29	3.06	3.00	3.46	3.18	2.85	3.31	3.48	2.49	4.23	2.23	3.11	3.54	3.71	5.82	5.22
š	2.96	5.04	4.87	3.90	4.54	5.22	3.78	3.01		5.34	5.34	3.49	4.50	3.71	4.16	4.57	4.19	4.04	2.92	3.94	3.87	3.53	2.91	3.67	4.87	4.74		5.34
t	2.26	4.11	4.28	5.22	5.12	4.23		2.52	4.68	3.68	3.02	1.39	3.39	3.36	3.57	3.95	2.80	2.60	3.55	2.10	3.29	4.71	2.66	2.87	3.23	2.50	5.82	4.46
u	2.80	3.80	4.00	3.59	3.86	3.59	4.54	3.22	4.74	4.16	4.28	2.39	3.42	3.71	3.11	3.27	3.13	3.05	3.54	3.42	2.86	4.04	2.86	2.80	3.86	3.24	4.18	4.29
v	2.31	3.37	3.39	4.37	5.22	4.02	4.57	2.91	3.77	3.64	4.07	1.28	2.94	3.82	3.32	4.11	2.45	2.74	4.06	3.32	4.25	5.12	3.75	3.37	3.20	4.74	5.22	
z	2.57	3.63	4.54	4.87		3.39		3.12	4.87	3.88	4.13	2.93	4.04	4.50	3.81	3.85	3.33	3.05	4.68	3.82	3.76	5.82	3.50	2.97	3.70	3.30	5.34	5.82
ž	3.18	4.82	5.82	5.12	3.34	4.74	4.16	3.13		5.82		3.00	4.30	5.34	4.16	4.44	3.79	4.36		4.78	5.22	5.22	4.34	4.07	4.48	5.04	3.86	
␣																												
@	2.11	3.22	3.71	4.00	4.16	3.37	4.48	2.42	3.92	3.72	3.69	2.55	3.35	3.54	3.57	3.22	3.07	2.42	2.97	3.43	2.68	2.92	3.06	2.44	3.57	3.26	4.30	4.36

Note: Rows and columns for letters with empty cells are omitted from the table.

Table 7. *"deletionCondOnFollowing"*—relative frequencies of errors where YX was mistyped as X (YX → X).

	a	b	c	č	ć	d	đ	e	f	g	h	i	j	k	l	m	n	o	p	r	s	š	t	u	v	z	ž	␣
a		3.40	3.51	3.51	3.89	2.89	4.26	2.73	4.72	3.46	3.19	3.59	2.17	2.74	2.38	2.96	2.44	3.64	3.08	2.16	3.28	3.74	2.38	3.02	2.68	3.36	3.63	
b	3.16					3.31		2.48			4.27	3.71	3.46	5.35	4.28	3.30	4.87	2.74	5.65	3.76	5.00		4.95	3.81		3.78	3.80	1.94
c	2.54			4.57		4.00		3.31			5.47	2.37	4.12	3.04	3.89	5.95	2.48	3.64	3.78	3.71	3.27	5.95	3.91	3.69	4.11			4.91
č	2.91					5.17		3.49				2.77	3.69	4.80	5.17	4.49	3.65	3.41	4.87	4.07		4.54		3.26	4.70			4.36
ć	3.79							3.27				3.64		3.94		5.35		3.95	3.87	4.87		3.48		3.34				
d	2.48	5.65						2.35		4.17	5.47	3.18	3.86		3.85		2.62	2.60	5.11	3.21	5.17		4.70	3.34	4.50	3.20	5.95	2.74
đ	3.55							3.53				4.50					4.84	3.77		3.81				4.54			5.17	
e	3.26	3.40	3.21	3.47	3.74	3.12	3.87		3.59	3.30	3.51	3.42	1.34	2.86	2.25	2.90	2.53	4.32	3.27	2.28	3.21	3.89	2.50	4.24	2.98	3.83	3.50	
f	3.50					4.75		3.55			5.65	3.65	5.17	5.05	4.26	4.84	3.45	3.70	5.00	4.13	3.35		3.89	4.95		5.95		4.08
g	2.93					3.49		3.07	5.95		5.95	2.41	4.34	5.17	4.10	5.65	2.86	2.57		3.33	5.95		4.80	3.40		3.32		3.89
h	3.57		3.22			4.72		3.95	5.65	4.45		2.68	5.05	4.95	4.80		4.21	4.15	4.59	3.47	3.93		3.70	4.34		4.75		5.65
i	3.41	3.43	2.87	3.25	3.64	2.75	4.61	3.09	3.57	3.75	3.10		2.12	2.82	2.20	3.06	2.05	3.00	2.98	2.24	2.78	3.82	2.14	3.88	2.90	3.14	3.60	3.40
j	2.49	3.69	4.24	4.67		3.35		3.37				0.62		5.95	2.52	3.18	2.53	2.77	3.62	3.93	3.14		3.38	3.00	3.18	5.17	5.35	3.81
k	2.57	4.70	3.74	3.36	5.47	5.95		2.77		4.41	4.84	2.26	3.41		3.81	4.72	2.72	2.90	3.96	3.46	2.36	3.60	2.36	3.44	4.16			2.90
l	2.27	3.48	4.91	4.55	5.47	3.29		2.56	4.28	3.32	4.29	2.51	3.62	2.75		3.45	5.05	2.35	2.85	3.59	2.86	4.00	3.65	3.25	3.12	3.17	4.84	3.05
m	2.54					3.91		2.45	5.95	4.80	4.46	2.47	3.44	3.78	3.54		3.19	2.55	5.47	2.83	3.40	5.35	3.62	3.08	5.25	3.34	4.80	1.92
n	1.98	3.25	5.47	3.26	3.97	2.31		1.85	5.25	3.97	3.68	2.16	2.76	3.63	2.58	3.48		2.53	3.56	2.68	2.50	3.55	2.53	3.16	2.86	3.24	3.91	4.01
o	2.17	3.55	3.55	4.39	4.80	2.98	5.35	2.72	3.64	3.25	3.12	2.55	2.50	2.38	2.53	3.01	2.01		2.69	2.28	3.42	5.17	2.68	3.30	2.68	3.66	5.95	3.66
p	3.00					5.05		3.26			5.47	3.54	3.41	5.11	4.45	2.89	4.95	2.80		3.22	2.47	5.35	3.07	3.24	5.95	5.65		3.28
r	2.35	3.53	3.91			2.98		2.32	4.43	3.31	3.81	2.29	4.13	3.15		4.42	3.76	2.15	2.55		3.44		2.65	2.87	3.05	3.65	5.25	3.70
s	2.74	4.52	4.95			2.51		2.44	4.24	4.43	4.32	2.46	2.31	3.19	3.65	3.91	2.29	2.56	3.67	3.07			2.40	3.19	2.91	5.65		2.71
š	3.20					4.10		3.31			4.80	3.10	4.26	3.96			3.93	3.22	3.85	2.87			5.35	3.58	3.76			4.45
t	2.18		3.78			4.31		2.42	4.24	5.65	3.41	2.23	3.62	2.74	3.16	5.65	2.01	2.86	3.37	2.86	1.89	3.30		2.84				2.92
u	3.17	3.64	3.74	3.87	4.06	3.48	4.22	3.01	4.67	3.67	3.97	4.30	2.34	2.98	2.93	3.20	2.74	2.95	3.21	2.54	3.35	5.11	2.72		3.07	3.52	4.45	4.23
v	2.17	4.46	5.65	5.65		3.46		2.92		5.17	3.17	2.46	3.56	2.75	4.95	5.95	3.58	2.28		3.17	3.19	5.95	2.62	3.72		3.09	5.05	2.76
z	2.79	4.28				3.96		3.34		5.47		2.53	3.96		4.87		2.53	2.69	5.65	3.46	5.35		4.78	3.30				2.07
ž	3.25					3.37		3.67				3.76	5.65				4.20	3.73		3.04				3.67				3.42
␣	1.99											1.41				4.84					5.95							
@	1.77	4.75	3.90	4.46	4.11	3.17		2.02	4.28	3.24	2.88	1.88	2.84	3.15	3.35	2.47	2.42	2.37	3.87	2.57	2.97	5.65	2.41	2.70	3.35	4.33		

Note: Rows and columns for letters with empty cells are omitted from the table.

As expected, the deletion matrix conditioned on the removal of letter Y in front of letter X reveals the common error of the wrong usage of "ije" and "je", where "i" was removed from the proper form. This error accounts for 23.73% of all spelling errors in the dataset.

This matrix also shows cases where "j" was removed in before "e", which happens mostly when texts in Serbian come for processing and use words that in Croatian contain "je" but in Serbian are written without "j" (e.g., "ponedjeljak" [Monday] is written as

"ponedeljak", "gdje" [where] as "gde", or "čovjek" [human] as "čovek"). Since this error falls under an edit distance of 1, corrections to proper Croatian forms are offered. This particular error accounts for 4.57% of all errors.

Another error that is visible in this matrix (3.9% of all errors) is removal of the letter "i" in front of a space, which often happens when the infinitive of the verb is used in its shortened form—e.g., "ponoviti UZV" [to repeat the ultrasound] is spelled as "ponovit UZV".

5.4. "deletionCondOnPrevious" Matrix

Table 8 shows the relative frequencies of errors where XY was erroneously written as X (XY → X). It is not that obvious to find the winner here, but upon closer examination, it is noticeable that the letter "i" (represented by column "i") deleted after "d", "r", "v", "l", or "m" (represented in their rows) has greater frequency, which is actually a consequence of removing "i" before "j", where letters "d", "r", "v", "l", or "m" should stand before "i". To illustrate this, "primjetiti" should be "primijetiti" [to notice], "poslje" should be "poslije" [after], and "djete" should be "dijete" [child]. This clearly illustrates the need for the deletionCondOnFollowing matrix, where all these examples would fall under one mistake, deleting "i" before "j".

Table 8. *"deletionCondOnPrevious"*—relative frequencies of errors where XY was mistyped as X (XY → X).

	a	b	c	č	ć	d	đ	e	f	g	h	i	j	k	l	m	n	o	p	r	s	š	t	u	v	z	ž	␣
a	4.95	3.87	3.27	3.35	4.31	2.60	4.05	3.93	3.98	3.58	2.90	3.62	2.33	2.25	2.41	3.09	1.94	3.07	3.22	2.67	2.46	3.58	2.48	3.25	2.51	2.67	3.76	2.40
b	2.86	4.23				4.95		3.42		5.35	5.25	2.17	2.76		3.08	4.80	3.03	3.03		3.07	4.91			3.84	4.17	4.20		
c	2.89		3.22					2.85			3.38	1.81	3.04	3.53	4.65		5.47	3.83	5.95	3.83	5.11		3.60	3.88	3.93	5.95		
č	3.02		4.84					2.67				2.71	3.64	2.80	3.98	5.65	3.17	4.84		5.95				3.33	5.17			
ć	3.49							3.34				2.66	5.95	4.87			3.79	5.35						4.21				
d	2.42	3.51	4.84			3.12		2.48		3.68	4.78	1.56	2.13		3.50	4.17	2.59	2.75		2.42	2.93	5.17	4.78	3.09	3.57	4.67	3.54	5.65
đ	3.73							3.41				4.30					5.65	5.17						3.95				
e	2.55	3.20	3.22	3.59	3.66	2.18	3.99	3.44	3.99	3.64	3.60	3.43	3.24	2.69	2.61	2.90	1.87	2.64	3.37	2.21	2.40	3.46	2.46	3.43	3.38	3.29	4.00	1.43
f	4.13							3.50	3.48			3.19		4.84	4.24		5.47	3.66		3.31	4.14		3.65	4.72				
g	2.67					4.16		3.02	5.95	3.22	3.40	3.38	5.65	5.35	2.31	4.54	3.62	2.68		2.86	5.47		5.11	2.93	4.80	4.75		
h	3.60	5.95	5.95					3.81			4.95	3.49		4.57	4.65	5.00	3.53	3.10		3.67	4.57		3.57	4.35	3.81			
i	3.85	4.47	3.26	3.11	3.98	3.28	5.11	3.23	4.30	3.71	3.73	4.50	1.89	2.59	2.87	2.93	2.21	2.68	3.34	2.99	2.28	3.32	2.59	4.75	2.95	2.82	4.33	3.86
j	2.21	4.24	4.80	4.78		4.67		1.91	5.65	4.61		2.23	3.09	3.96	4.54	4.45	2.95	3.69	4.91	5.47	3.01	4.52	3.92	2.96	4.57	4.95		5.95
k	2.37	5.95	3.76	5.11	5.65	5.47		3.33	5.00	4.72	4.41	3.08	5.35	5.17	2.80	4.11	3.76	2.39	5.95	2.88	2.89	5.05	2.58	3.20	2.96			
l	2.28	5.05	5.05	5.65		4.36		2.66	5.35	5.17	5.47	1.41	2.03	4.33	2.77	4.46	2.76	2.64	4.80	5.95	4.42		3.10	3.36	4.59	4.75		
m	2.08	3.67	5.35	4.75	5.95	5.65		2.44	4.95			1.11	2.12	5.00	3.44	2.83	3.02	2.79	2.78	3.89	3.85		5.11	3.70	5.00			4.75
n	2.24	5.35	3.36	4.42		3.11	5.95	2.30	3.93	3.24	4.63	1.56	1.87	3.01	4.70	5.65	2.85	2.43	5.65	5.95	2.72	5.95	2.49	2.84	4.16	3.67	5.47	
o	3.38	3.06	3.80	3.86	4.46	2.49	4.30	3.80	4.15	3.30	3.91	2.92	2.74	3.17	2.31	2.36	2.25	2.72	2.79	2.46	2.29	3.37	2.50	3.24	2.86	3.36	3.93	2.98
p	2.64	5.65	4.57	5.47	4.59	5.95		3.11	5.35		4.41	2.31	2.94	4.63	2.93	5.11	3.70	2.25	2.86	2.31	3.73	4.80	3.45	3.37				
r	1.84	4.50	3.76	4.78	5.05	2.96	4.80	1.83	4.87	3.70	4.03	1.55	2.93	3.70	4.46	3.38	2.88	2.25	4.30	2.98	2.89	3.82	2.68	2.91	3.39	3.92	3.65	
s	1.89	5.95	3.30					2.48	4.01		3.16	2.43	2.40	2.63	2.76	3.53	2.71	2.86	3.00	3.83	2.63		1.93	3.27	3.39	5.95		3.79
š	2.99		5.25	4.31	3.87			3.24				3.07	5.17	3.42	3.73		3.52	4.16	4.08				2.66	4.87	5.25			5.35
t	2.12	5.65	4.57			4.75		2.26	4.27	5.95	3.10	1.18	2.76	3.13	3.76	4.42	2.41	2.60	3.35	2.04	2.91		3.06	2.77	2.47	5.11		
u	2.89	3.56	4.27	3.26	3.60	3.18	5.17	4.17	5.25	3.68	3.93	3.25	3.05	3.24	2.98	3.33	2.75	3.47	2.95	2.88	2.51	3.68	3.02	3.41	3.86	3.31	3.93	3.55
v	2.15	5.95	4.42	4.70		4.23		2.60		5.95		1.52	2.10	4.55	3.05		2.52	2.20		2.51	4.01	5.05		3.72				
z	2.51	3.73				3.17		3.33		3.37	5.11	2.78	4.29		3.38	3.52	2.71	3.39		3.51	5.00			3.10	3.24	4.78		3.16
ž	3.37	3.60				3.61	5.95	3.28		5.65		3.29	4.20		4.23	4.28	3.65	4.70		4.61				4.57	5.00			
␣						5.95																						
@	3.26	3.80	4.10	3.81	5.65	3.13	5.47	3.69	4.11	3.73	2.96	2.79	3.80	3.16	3.55	3.17	3.01	2.78	2.64	3.48	2.84	4.31	2.60	3.01	3.55	3.47	4.00	

Note: Rows and columns for letters with empty cells are omitted from the table.

The observations about the main diagonal in the insertion matrices are valid here as well. Even though two duplicate consecutive letters are not characteristic for Croatian, certain compound words feature them—e.g., "preddiplomski" [undergraduate], "najjači" [the strongest] or "samoobrana" [self-defense]. The main diagonal of the deletionCondOnFollowing matrix is empty because when the letter is erroneously missing (e.g., "samobrana"), the second letter "o" is considered missing and is accounted for in the intersection of row "b" and column "o". In deletionCondOnPrevious, it is counted in the intersection of row "o" and column "o", as it is treated as "o" missing after "o". However, this kind of error is negligible across the whole dataset because words with duplicate characters are far less frequent than others.

5.5. "Substitution" Matrix

Table 9 gives insight into the relative frequencies of errors where X was mistyped as Y (X → Y).

Table 9. "*Substitution*"—relative frequencies of errors where X was mistyped as Y (X → Y).

	a	b	c	č	ć	d	đ	e	f	g	h	i	j	k	l	m	n	o	p	r	s	š	t	u	v	z	ž
a		4.72	4.53	5.22	6.06	4.34	5.59	1.74	4.98	4.78	4.89	2.47	4.32	4.53	4.20	4.41	4.24	2.33	4.78	4.16	2.84	5.59	3.94	3.00	4.68	4.53	5.76
b	6.06		4.68	5.06	5.36	3.55	5.28	5.76	5.02	3.70	4.48		3.92	4.27	4.52	3.88	3.19	5.06	3.17	4.63	4.70	5.76	4.63	5.22	3.00	4.72	5.76
c	5.11	5.02		3.58	3.81	3.69	5.22	4.27	4.63	4.42	4.52	5.02	4.30	3.03	4.56	4.98	3.96	5.06	4.78	4.35	2.74	5.76	3.38	5.02	3.03	3.52	5.76
č	5.59	5.59	0.84		1.05	4.01	3.37	3.83	5.28	5.46	4.92	4.62	4.38	3.59	2.88	5.76	4.65	5.22	4.15	5.22	4.14	3.18	4.11	6.06	4.49	4.35	3.73
ć	4.62	5.76	1.05	0.79		3.51	4.12	4.03	4.86	5.02	5.76	5.36	4.49	5.22	4.26	4.98	4.72	6.06	5.06	5.46	4.57	4.13	3.16	5.28	4.74	5.76	3.28
d	4.70	3.55	3.67	4.65	4.95		3.35	4.28	3.53	3.60	4.83	5.06	4.32	4.07	4.20	4.78	3.55	4.81	4.09	3.49	2.77	5.36	2.53	5.22	3.94	3.89	4.98
đ	5.36	5.46	4.04	3.97	3.99	1.70		5.59	5.76	4.29	5.59		4.45			6.06	3.80		5.28		4.06	2.95	6.06		6.06	4.78	3.22
e	1.74	5.36	3.85	5.46	5.06	3.84			4.23	4.81	4.56	2.35	3.51	4.70	4.52	4.43	4.25	2.45	4.52	2.99	3.49	5.59	4.18	3.15	4.98	5.02	6.06
f	5.76	4.78	4.48	6.06		3.82		4.81		3.80	4.81	5.06	5.46	4.76	4.95	4.92	5.06	4.92	4.42	4.26	4.92	5.59	4.36	6.06	3.35	5.02	5.46
g	4.89	3.60	4.31	4.62	4.98	3.25	3.96	4.44	3.49		3.04	5.16	4.12	3.03	4.92	4.92	4.04	4.68	4.76	4.14	4.24	5.46	3.72	5.59	4.10	3.84	5.02
h	4.95	4.41	4.65	5.59	5.36	4.60		4.70	4.62	3.17		5.36	3.33	3.36	5.22	4.60	3.51	4.68	5.22	4.45	4.30	4.54	4.00	4.36	3.38	4.48	5.59
i	2.43	6.06	4.95	5.59	5.46	3.93		2.33	5.11	5.16	4.30		3.15	3.94	3.43	4.65	3.80	2.03	4.86	4.15	4.19	5.28	3.84	2.24	4.53	4.62	5.11
j	4.51	4.78	4.23	4.49	4.31	4.35	4.47	4.42	5.46	4.51	3.45	3.01		3.17	3.57	3.59	3.25	4.70	4.57	3.90	4.27	5.06	4.11	3.81	2.35	4.39	5.02
k	4.26	4.70	2.63	4.05	4.66	3.99	5.22	4.78	4.62	3.22	2.93	4.05	3.25		2.93	3.44	3.19	4.16	3.96	4.08	4.37	4.95	3.50	4.36	4.44	4.68	
l	4.57	4.72	4.49	3.02	4.49	4.05	5.76	4.33	4.52	4.40	5.16	3.11	3.46	2.97		3.90	3.05	3.50	3.74	3.37	3.88	4.89	3.68	4.34	4.04	4.72	4.86
m	4.15	3.84	4.52	5.28	5.36	4.21	5.46	4.66	5.46	4.20	4.24	4.76	3.52	3.18	3.80		2.30	4.52	3.79	4.16	4.33	4.78	4.20	4.74	3.79	4.81	5.59
n	4.01	3.09	3.77	4.86	4.74	3.60		4.19	5.28	3.82	3.68	3.99	2.99	3.28	2.79	2.32		4.35	4.27	3.52	3.98	5.06	3.30	4.14	3.47	3.90	
o	1.96	4.81	4.57	5.46		4.39	6.06	2.59	5.59	4.89	5.11	2.03	4.29	3.70	2.65	4.54	4.19		2.76	4.09	4.03	5.02	4.46	2.40	4.89	5.16	6.06
p	4.72	2.66	4.95	4.25	5.59	4.05	5.59	5.22	4.47	4.70	5.11	4.46	4.86	3.98	3.58	3.61	4.28	2.97		4.31	4.41	3.22	4.03	5.02	4.35	5.11	5.76
r	4.52	4.89	4.20	4.19	4.92	3.38	5.46	3.23	4.26	4.07	4.89	3.99	3.76	4.16	3.27	4.18	3.42	4.42	4.02		2.64	4.57	2.45	4.51	3.85	3.83	5.16
s	3.03	4.57	3.12	5.36	6.06	2.78	5.06	3.34	4.76	4.35	4.24	3.89	4.17	4.66	4.36	4.41	4.20	3.80	4.43	3.79		2.68	3.78	4.38	4.39	2.59	5.22
š	5.22	6.06	4.36	3.20	3.71	4.46	3.23	5.02	6.06	4.17	4.76	4.68	4.83	5.06	4.78	5.28	5.36	5.02	2.53	4.42	0.88		4.05	5.59	5.22	3.77	3.19
t	4.25	4.57	3.43	4.00	3.03	2.00	6.06	4.09	4.02	3.97	3.82	3.87	4.05	3.52	3.46	4.46	3.41	4.29	4.01	2.52	3.63	4.70		4.25	4.05	2.84	5.59
u	2.59	5.46	5.76	5.06	6.06	5.16		3.20	5.36	5.46	4.28	2.15	4.09	5.02	4.76	4.95	4.86	2.35	5.28	4.53	4.30	5.76	4.68		2.68	3.51	
v	4.86	2.88	3.00	4.74	5.22	4.01		5.36	3.35	4.11	4.14	4.38	2.55	4.19	4.11	3.67	3.49	3.92	4.46	4.31	4.28	6.06	4.00	4.59		4.28	5.59
z	4.89	5.46	3.64	4.76	5.16	4.19	6.06	4.86	5.11	3.52	4.54	4.95	4.74	4.60	4.53	5.22	4.12	4.81	5.22	3.91	2.68	5.16	2.93	3.52	4.78		3.08
ž	5.76	5.76	4.38	3.52	2.94	4.24	3.37	4.92	4.98	4.06	5.59	3.91	3.63		4.07			5.76		3.81	4.60	3.22	4.95	5.36	5.46	1.03	

Note: Rows and columns for letters with empty cells are omitted from the table.

Here, writing "č" instead of "ć" is the most common error—it happens in 16% of all substitutions (row "ć", column "č"), with most notable examples being "mogučnost" instead of "mogućnost" [possibility] and "čemo" instead of "ćemo" [we will]. However, writing "ć" instead of "č" happens half as often (row "č", column "ć"), e.g., "naćin" instead of "način" [way, method] and "inaće" instead of "inače" [otherwise]. Also, this matrix shows that often both "ć" and "č" are substituted with "c", "đ" with "d" (but less often, as "đ" is not a frequent character"), "š" is substituted with "s", and "ž" with "z". Substituting "đ" with "dj" produces an error of Damerau–Levenshtein distance 2 (one substitution and one insertion) and is not accounted for in this research. Substituting "dž" with "dz" is also common (even though "dž" is even less frequent than "đ") but is accounted for in the substitution of "ž" with "z" already because the data was analyzed at character level. Another spelling error that can be observed from the data is related to assimilation of consonants. The substitution of "t" with letter "d" (ranked 11th, with a relative frequency of 0.010049) in examples such as "predpostavljam" (proper form: "pretpostavljam" [I assume]) or "predhodno" (proper form: "prethodno" [previous]) is a consequence of users' unawareness of the assimilation rule, where "d" in front of "p" should become "t". Other errors are also observable but are not prominent (e.g., "pretstavlja" instead of the proper "predstavlja" [presents], "sretstva" instead of the proper "sredstva" [means, resources], "substanca" instead of the proper "supstanca" [substance], "drugčije" instead of the proper "drukčije" [differently], etc.).

5.6. "Transposition" Matrix

Table 10 shows the relative frequencies of errors where adjacent letters XY were misspelled as YX (XY → YX).

Table 10. *"Transposition"*—relative frequencies of errors where XY was mistyped as Y (XY → YX).

	a	b	c	č	ć	d	đ	e	f	g	h	i	j	k	l	m	n	o	p	r	s	š	t	u	v	z	ž
a		3.36	3.16	3.12	5.18	2.59	5.00	3.28	4.18	3.10	3.27	3.43	2.03	2.01	1.63	1.97	1.86	2.50	2.46	1.85	2.71	3.68	2.00	3.45	1.97	2.76	3.40
b	3.07					3.82		3.24				3.08	4.70		3.23	4.87	3.62	2.66		3.48				3.92	4.87	3.22		
c	3.04							3.60			3.98	2.24	4.52	4.11	4.70			3.61		3.78	4.22		3.90	3.50	5.48	5.00	
č	2.91							3.28				3.70		4.01	5.48		3.78							4.15			
ć	4.20							3.55				4.27															
d	2.65	3.90						3.35		3.64		2.12	2.59		3.15	3.37	1.72	2.31		3.05	3.64		5.48	3.07	3.17	4.03	3.96
đ	4.40							4.36																5.18			
e	3.66	3.08	3.78	3.82	3.72	2.67	4.01		4.20	3.55	3.61	2.55	3.18	2.13	2.17	2.32	1.71	2.53	3.08	1.81	2.76	3.75	2.48	3.55	3.32	3.71	3.89
f	4.70							2.40				3.62			3.93		4.87	3.02		4.27	5.00		4.63	4.52			
g	2.90					3.92		3.39			4.03	3.51		5.48	2.78	4.87	3.14	2.62		3.26	5.00			3.25		4.57	
h	3.55	5.48						3.93				4.06		4.78	5.48	5.48	4.15	3.91	5.48	4.06	5.00		3.31	4.40	3.44		
i	3.37	4.43	2.44	3.46	4.63	3.00		3.26	3.84	3.35	3.29		2.54	3.17	2.40	3.04	2.14	3.39	3.28	2.36	1.91	3.56	2.00	4.10	2.61	2.43	5.00
j	1.57	5.00	4.70	5.48		3.77		1.15				2.46		4.15	4.04	3.70	2.78	3.51	5.48	4.70	2.69		3.59	3.68	3.16	4.78	
k	1.94		2.93	4.70				2.36			5.18	3.43	5.48		3.20	4.57	3.69	3.20	4.63	2.56	2.45	3.94	2.59	3.75	3.34		
l	1.51	4.11	4.78	5.48		3.93		1.76	4.57	4.36	5.18	2.53	2.68	3.61		3.98	3.31	2.72	4.00	4.48	3.45		3.28	3.08	4.48	4.63	
m	2.00	3.51		5.18		4.57		1.98	5.00			2.90	3.23	5.18	4.78		3.98	3.16	3.25	4.25	3.19			3.31	5.18		
n	1.72		3.06	3.92		2.58	5.48	1.76	4.15	3.35	4.04	2.33	2.70	3.10	4.52	5.48		2.48		4.22	2.21	5.00	2.32	3.03	4.36	3.29	5.18
o	3.12	2.64	3.90	4.78	5.48	2.28		3.93	3.72	2.36	4.70	3.35	3.43	2.93	2.82	2.71	2.68		2.73	1.88	1.79	4.33	3.07	3.54	2.16	3.29	4.70
p	2.55		3.96					2.91	5.48		4.57	3.30	4.15	4.78	3.56	4.87	4.43	2.75		2.25	2.99	4.78	3.31	3.26			
r	1.28	3.80	3.96			3.25	5.00	2.05	3.90	2.75	3.76	2.04	4.06	3.07	3.75	2.66	2.95	1.43	3.21		2.98	3.42	2.77	2.21	3.42	3.70	3.92
s	3.21		3.94			5.48		3.23	4.78	5.48	3.80	2.56	2.90	2.24	2.39	3.17	2.02	2.68	2.51	4.52			2.42	2.87	3.82	4.63	
š	3.06			5.00	4.20			3.41				3.44		4.22	5.48			4.63					3.38	4.52			
t	2.02		4.27			5.48		2.27	4.20		3.01	1.70	2.96	2.67	3.77	3.72	2.12	2.17	2.69	2.71	2.57			2.35	3.42	4.87	
u	3.06	3.71	3.85	3.96	4.30	3.17		4.33	4.78	3.60	5.18	3.85	3.27	3.40	2.73	3.50	3.35	3.21	2.95	2.61	2.39	4.63	2.30		3.74	3.49	4.52
v	1.70	5.48	4.52			4.27		1.74				2.50	3.10	3.78	3.09		2.60	1.78		2.75	4.70	4.15		3.77			
z	2.93	3.82				3.29		3.55		4.78	4.87	2.91	5.18		4.00	3.87	3.27	3.49		3.52			5.00	3.65	3.67		
ž	3.21	3.69				4.33	4.87	3.64				4.57	4.87				5.48								5.00		

Note: Rows and columns for letters with empty cells are omitted from the table.

Unlike in other presented confusion matrices, in this case, the deviations from random typos were not observed. Even though some errors dominate, compared to other types of errors, they show a more uniform distribution where even proximity of keys on the keyboard does not contribute much to the error.

It seems that the letter "a" is transposed more frequently, either with a group of letters that are usually typed with the right hand or adjacent letters typed with the left hand. For example, "pozdarv" is often written instead of "pozdrav" [greeting] (row "r", column "a") and "stavri" instead of "stvari" [things] (row "v", column "a").

This may lead to the conclusion that different speeds at which the left and right hands work can have a notable impact on the correct spelling of the written text. In cases where there is a significant imbalance in typing speed between the two hands or even between two fingers of one hand, errors can occur because one hand or finger is faster than the other. This discrepancy can lead to typos, misspellings, or even omitted letters, as the faster hand may accidentally skip characters or anticipate the next ones before they have been typed correctly.

Disparity in typing speed becomes even more noticeable when typing fast and can potentially compromise the overall accuracy of the written content. This emphasizes the importance of refining typing skills and maintaining a harmonious balance between the left and right hands to improve typing and spelling and, subsequently, produce error-free text.

5.7. Implementation of Confusion Matrices in Spellchecking

As a proof of concept, we used our matrices in the process of sorting correction candidate words in a list of possible corrections offered to the user. After selecting all the possible correction candidates with edit distance 1, we sorted the correction candidates based on the product of the relative frequency of the correction candidate word from our unigram corpus and the relative frequency of a given type of error that could convert the correct word to the wrong word. For example, given the error word "prjetili", the only two correction candidates are "prijetili" [threatened], and "pretili" [obese]:

- The relative frequency of the word "prijetili" in our corpus is 2.4966×10^{-6}. In order to mistype "prijetili" as "prjetili", a deletion of character "i" in front of "j" is required, and according to Table 7, *deletionCondOnFollowing* at [40] (row "j", column "i"), the relative frequency of such an event is 0.237323. The product of those two values is 5.925×10^{-7}.

- The relative frequency of the word "pretili" is 6.85977×10^{-7}. For mistyping "pretili" to "prjetili", we need to find the relative frequency of "j" inserted before "e" in Table 5 *insertionCondOnFollowing*—it is 0.007118. The product of the two values is 4.88278×10^{-9}.

Therefore, the word "prijetili" is offered as the first choice. However, the sentence, "Naši susjedi su prjetili." could be either "Naši susjedi su prijetili." [Our neighbors threatened.] or "Naši susjedi su pretili." [Our neighbors are obese.], which clearly shows the need to take into account the context. Although initial results of the implementation of our matrices show promising results, this research is still ongoing.

5.8. Log Charts

When spelling errors occur, users are more likely to label them as typos than to admit their poor knowledge of the orthography (i.e., spelling) rules. The difference is clear: if someone writes "adn" or "teh" instead of "and" or "the", it is a typo. However, if a person writes "than" instead of "then" or "wellcome" instead of "welcome" it may be assumed it is not a typo but a sign of unfamiliarity with orthography rules.

It is possible to use the data from the presented matrices to visualize the relative frequencies of the errors on a logarithmic scale and try to determine which of the explanations for the error is more likely: is it a typo, is it a lack of proficiency in orthography, or are users simply saving time by replacing a letter with a simpler variation that requires fewer keystrokes?

The relative frequencies of spelling errors from the confusion matrices are shown graphically in Figure 2a–f according to the principle of rank-size distribution in decreasing order of size. The rank of each error type is shown on the x-axis, and the corresponding relative frequency is shown on the y-axis. Due to the large range of magnitudes, the values on both axes are on a logarithmic scale in order to make their dependence visible.

This way of visualizing data corresponds to the Zipf–Mandelbrot distribution [41], an empirical law that is often used for describing linguistic phenomena, e.g., in a certain language, the frequency of each word is inversely proportional to its rank in the frequency distribution.

As can be seen from Figure 2a–e, the points corresponding to the higher ranks are distributed as if forming a smooth and regular curve, while for lower ranks, the values of the points may deviate significantly upwards from the supposed curve. However, in the case of a transposition spelling error, as shown in Figure 2f, points of a lower rank do not have a specified observed deviation. This fact confirms that transposition errors are random in nature.

In all other cases, there are individual errors that deviate significantly from randomness and are marked by red dots in Figure 2a–e. Such an approach could be used to identify spelling error outliers, i.e., extremely frequent errors, as explained in the discussion section. In future related research, modeling of the curve will be performed, so the level of deviation from the curve will enable an objective quantitative judgment of what is a spelling error and what is due to ignorance.

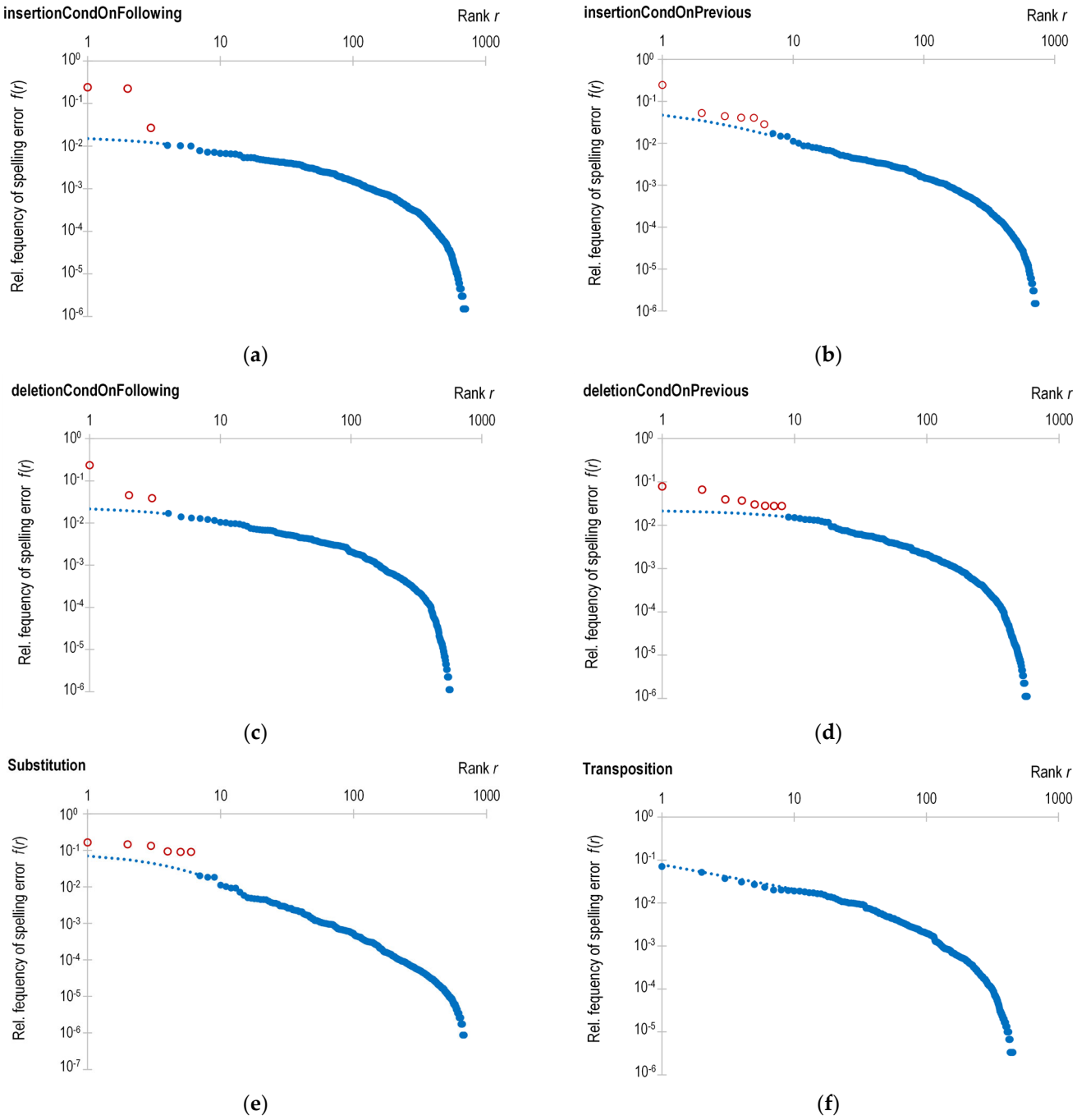

Figure 2. Log–log plot of relative frequencies of spelling errors for the (**a**) insertion of letter Y conditioned on the previous character, (**b**) insertion of letter Y conditioned on the following character, (**c**) deletion of letter Y conditioned on the previous character, (**d**) deletion of letter Y conditioned on the following character, (**e**) substitution where X was mistyped, and (**f**) transposition where adjacent letters XY were mistyped as YX. The red dots represent deviations from the rank-size distribution expected trend (indicated by the blue dashed curve).

6. Conclusions

Spellcheckers are indispensable tools in the current digital age, both for everyday writing and for professional communication. They can quickly identify and correct spelling errors, improving the readability and quality of texts, especially for non-native speakers. These tools are more than just error correctors. They also help users to improve their

language skills. In the professional field, e.g., for academic papers or legal documents, the accuracy of spellchecking is essential.

Our research, based on an experimental dataset derived from a long-term collection of mistyped words and user corrections, presents a novel approach to leveraging confusion matrices for spellchecking error pattern discovery and the improvement of spellchecker precision in the Croatian language. Our findings contribute to the advancement of Croatian spellchecking technologies, particularly in providing a more accurate offering of correction candidates. Our work offers a deeper understanding of linguistic specifics, particularly in underresourced languages with rich orthographies like Croatian.

The study has uncovered subtle statistical properties of spelling errors in the Croatian language, emphasizing the development of spellcheckers and the crucial role of confusion matrices in refining suggested corrections. The user-generated data from the Croatian spellchecker ispravi.me has been examined to provide insights into common spelling errors which may be used for the creation of confusion matrices based on the linguistic details of the Croatian language.

The research conducted shows the importance of using user data to improve the accuracy of spellchecking algorithms. By examining the frequency and patterns of corrections, matrices were created that not only statistically evaluate the performance of current spellcheckers but also provide a basis for future improvements to these important digital tools on the web and mobile devices. The implications of the data obtained go beyond spellcheckers and provide a deeper understanding of the linguistic challenges posed by the use of diacritics and the accessibility of virtual keyboards.

Concerning future development, the user-driven confusion matrices presented in this paper pave the way for further advances in the field of spellchecking, especially in languages with unique orthographic features. The context-dependent nature of the presented approach opens new possibilities for more accurate and linguistically informed correction suggestions, thus contributing to the ongoing evolution of language processing tools.

Finally, it is important to emphasize the dynamic nature of language use and the need for adaptive technologies. Future research efforts could use the findings reported in this study to improve spellcheckers, investigating additional aspects of language data in order to improve the overall user experience in different linguistic contexts. Such a user-centric approach extends the scope of spellchecking and also emphasizes the importance of incorporating user data to customized language processing tools for achieving better performance and user satisfaction.

Author Contributions: Conceptualization, G.G. and M.H.; methodology, G.G. and M.S.; software, G.G. and M.S.; validation, M.M. and M.S.; formal analysis, M.M.; investigation, M.S. and G.G.; resources, G.G.; data curation, G.G. and M.M.; writing—original draft preparation, G.G.; writing—review and editing, M.S., M.H. and M.M.; visualization, M.S.; supervision, G.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset is described in [32]. The results obtained on a subset of the dataset which are presented in this paper are available in detail at https://ispravi.me/confusion (accessed on 31 December 2023).

Acknowledgments: We express our heartfelt gratitude to Šandor Dembitz, whose visionary leadership has been instrumental in advancing language technologies and spellchecking for the Croatian language. His outstanding contribution reflects not only a brilliant mind but also a commitment to innovation. We appreciate Dembitz's dedication to the development of the spellchecking service, which has significantly enhanced the language landscape for Croatian.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mitton, R. Fifty Years of Spellchecking. *Writ. Syst. Res.* **2010**, *2*, 1–7. [CrossRef]
2. Hládek, D.; Staš, J.; Pleva, M. Survey of Automatic Spelling Correction. *Electronics* **2020**, *9*, 1670. [CrossRef]
3. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st ed.; Prentice Hall PTR: Hoboken, NJ, USA, 2000.
4. Almutiri, T.; Nadeem, F. Markov Models Applications in Natural Language Processing: A Survey. *Int. J. Inf. Technol. Comput. Sci.* **2022**, *14*, 1–16. [CrossRef]
5. Pienaar, W.; Snyman, D.P. Spelling Checker-Based Language Identification for the Eleven Official South African Languages. In Proceedings of the Twenty-First Annual Symposium of the Pattern Recognition Association of South Africa, Stellenbosch, South Africa, 22–23 November 2010.
6. Abandah, G.; Suyyagh, A.; Khedher, M.Z. Correcting Arabic Soft Spelling Mistakes Using BiLSTM-Based Machine Learning. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 815–829. [CrossRef]
7. Zhang, D.; Li, Y.; Zhou, Q.; Ma, S.; Li, Y.; Cao, Y.; Zheng, H.-T. Contextual Similarity Is More Valuable Than Character Similarity: An Empirical Study for Chinese Spell Checking. In Proceedings of the ICASSP 2023—IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
8. Jin, W.; Zhao, B.; Zhang, Y.; Sun, G.; Yu, H. Fintech Key-Phrase: A New Chinese Financial High-Tech Dataset Accelerating Expression-Level Information Retrieval. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*, 1–37. [CrossRef]
9. Davlatova, M. Semantic Properties of Effective Constructions in English and Uzbek Languages. *E3S Web Conf.* **2023**, *420*, 10027. [CrossRef]
10. Phatak, S.A.; Lovitt, A.; Allen, J.B. Consonant Confusions in White Noise. *J. Acoust. Soc. Am.* **2008**, *124*, 1220–1233. [CrossRef] [PubMed]
11. Xu, D.; Wang, Y.; Metze, F. EM-Based Phoneme Confusion Matrix Generation for Low-Resource Spoken Term Detection. In Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, NV, USA, 7–10 December 2014; pp. 424–429.
12. Kernighan, M.D.; Church, K.W.; Gale, W.A. A Spelling Correction Program Based on a Noisy Channel Model. In Proceedings of the COLING 1990 Volume 2: Papers Presented to the 13th International Conference on Computational Linguistics, Helsinki, Finland, 20–25 August 1990.
13. Cekaite, A. Collaborative Corrections with Spelling Control: Digital Resources and Peer Assistance. *Int. J. Comput. Support. Collab. Learn.* **2009**, *4*, 319–341. [CrossRef]
14. Mossige, M.; Arendal, E.; Kongskov, L.; Svendsen, H.B. How Do Technologies Meet the Needs of the Writer with Dyslexia? An Examination of Functions Scaffolding the Transcription and Proofreading in Text Production Aimed towards Researchers and Practitioners in Education. *Dyslexia* **2023**, *29*, 408–425. [CrossRef] [PubMed]
15. META-NET White Paper Series Key Results and Cross-Language Comparison. Available online: http://www.meta-net.eu/whitepapers/overview (accessed on 12 April 2023).
16. Sussex, R.; Cubberley, P. *The Slavic Languages*; Cambridge University Press: Cambridge, UK, 2006; ISBN 9780521223157.
17. Golubović, J.; Gooskens, C. Mutual Intelligibility between West and South Slavic Languages. *Russ. Linguist.* **2015**, *39*, 351–373. [CrossRef]
18. Nouza, J.; Safarik, R.; Cerva, P. ASR for South Slavic Languages Developed in Almost Automated Way. In Proceedings of the Interspeech 2016, ISCA, San Francisco, CA, USA, 8–12 September 2016; pp. 3868–3872.
19. Pedrazzini, N.; Eckhoff, H.M. OldSlavNet: A Scalable Early Slavic Dependency Parser Trained on Modern Language Data. *Softw. Impacts* **2021**, *8*, 100063. [CrossRef]
20. Adamou, E.; Breu, W.; Scholze, L.; Shen, R.X. Borrowing and Contact Intensity: A Corpus-Driven Approach from Four Slavic Minority Languages. *J. Lang. Contact* **2016**, *9*, 513–542. [CrossRef]
21. Banasiak, D.; Mierzwa, J.; Sterna, A. Extended N-Gram Model for Analysis of Polish Texts. In *Man-Machine Interactions 5*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 355–364.
22. Ziolko, B.; Skurzok, D.; Michalska, M. Polish N-Grams and Their Correction Process. In Proceedings of the 2010 4th International Conference on Multimedia and Ubiquitous Engineering, IEEE, Cebu, Philippines, 11–13 August 2010; pp. 1–5.
23. Rozovskaya, A. Spelling Correction for Russian: A Comparative Study of Datasets and Methods. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Varna, Bulgaria, 6–8 September 2021; Mitkov, R., Angelova, G., Eds.; INCOMA Ltd.: Moscow, Russia, 2021; pp. 1206–1216.
24. Sorokin, A. Spelling Correction for Morphologically Rich Language: A Case Study of Russian. In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 4 April 2017; pp. 45–53.
25. Richter, M.; Stranak, P.; Rosen, A. Korektor—A System for Contextual Spell-Checking and Diacritics Completion. In Proceedings of the COLING 2012: Posters, Mumbai, India, 8–15 December 2012; Kay, M., Boitet, C., Eds.; The COLING 2012 Organizing Committee: Mumbai, India; pp. 1019–1028.
26. Ramasamy, L.; Rosen, A.; Stranák, P. Improvements to Korektor: A Case Study with Native and Non-Native Czech. In Proceedings of the ITAT 2015: Information Technologies—Applications and Theory, Slovensky Raj, Slovakia, 17–21 September 2015.
27. Hladek, D.; Stas, J.; Juhar, J. Unsupervised Spelling Correction for Slovak. *Adv. Electr. Electron. Eng.* **2013**, *11*, 2013. [CrossRef]

28. Stankevičius, L.; Lukoševičius, M.; Kapočiūtė-Dzikienė, J.; Briedienė, M.; Krilavičius, T. Correcting Diacritics and Typos with a ByT5 Transformer Model. *Appl. Sci.* **2022**, *12*, 2636. [CrossRef]
29. Náplava, J.; Straka, M.; Straková, J. Diacritics Restoration Using BERT with Analysis on Czech Language. *Prague Bull. Math. Linguist.* **2021**, *116*, 27–42. [CrossRef]
30. Dembitz, Š.; Gledec, G.; Sokele, M. An Economic Approach to Big Data in a Minority Language. *Procedia Comput. Sci.* **2014**, *35*, 427–436. [CrossRef]
31. Šoić, R.; Vuković, M. N-Gram Based Croatian Language Network: Application in a Smart Environment. *J. Commun. Softw. Syst.* **2022**, *18*, 63–71. [CrossRef]
32. Šantić, N.; Šnajder, J.; Dalbelo Bašić, B. Automatic Diacritics Restoration in Croatian Texts. In Proceedings of the 2nd International Conference The Future of Information Sciences (INFuture 2009), Zagreb, Croatia, 4–6 November 2009; pp. 309–318.
33. Dembitz, Š.; Gledec, G.; Randić, M. Spellchecker. In *Wiley Encyclopedia of Computer Science and Engineering*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2009.
34. Wikimedia Commons contributors Croatian Keyboard Layout 2010. Available online: https://commons.wikimedia.org/wiki/File:Croatian_keyboard_layout.jpg (accessed on 31 December 2003).
35. Institute for Croatian Language and Linguistics Hrvatski Pravopis (Croatian Ortography). Available online: http://pravopis.hr/ (accessed on 31 December 2023).
36. Gledec, G.; Horvat, M.; Mikuc, M.; Blašković, B. A Comprehensive Dataset of Spelling Errors and Users' Corrections in Croatian Language. *Data* **2023**, *8*, 89. [CrossRef]
37. Dembitz, Š.; Gledec, G.; Blašković, B. *Architecture of Hascheck—An Intelligent Spellchecker for Croatian Language*; LNAI; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6277, ISBN 3642153895.
38. Gledec, G.; Šoić, R.; Dembitz, Š. Dynamic N-Gram System Based on an Online Croatian Spellchecking Service. *IEEE Access* **2019**, *7*, 149988–149995. [CrossRef]
39. Damerau, F.J. A Technique for Computer Detection and Correction of Spelling Errors. *Commun. ACM* **1964**, *7*, 171–176. [CrossRef]
40. Srdić, I.; Gledec, G. Confusion Matrices for Croatian Language. Available online: https://ispravi.me/confusion/ (accessed on 31 December 2023). (In Croatian)
41. Mandelbrot, B. An Informational Theory of the Statistical Structure of Language. In *Communication Theory*; Academic Press: Princeton, NJ, USA, 1953; pp. 486–502.